Multivariate Robust MRCD Based Hotelling's T2 Control Chart with Bootstrap Control Limit for Intrusion Detection

Ichwanul Kahfi Prasetya , Muhammad Ahsan * , Muhammad Mashuri , Muhammad Hisyam Lee

Posted Date: 27 December 2023

doi: 10.20944/preprints202312.2062.v1

Keywords: bootstrap; intrusion detection; multivariate control chart; MRCD; Hotelling's T2

Article

Multivariate Robust MRCD Based Hotelling's T^2 Control Chart with Bootstrap Control Limit for Intrusion Detection

Ichwanul Kahfi Prasetya [1], Muhammad Ahsan [1,*], Muhammad Mashuri [1] and Muhammad Hisyam Lee[2]

[1] Department of Statistics, Institut Teknologi Sepuluh Nopember. Surabaya, Indonesia
[2] Department of Mathematical Sciences, Universiti Teknologi Malaysia, Johor Bahru, Malaysia; ichwankahfi@gmail.com (I.K.P.), m_mashuri@statistika.its.ac.id (M.M), mhl@utm.my (M.H.L.)
* Correspondence: muh.ahsan@its.ac.id (M.A.)

Abstract: Intrusion detection is generally carried out by matching network traffic patterns with known attack patterns or by identifying abnormal network traffic patterns. One statistical methodological approach used in intrusion detection is Statistical Process Control (SPC) by constructing a control chart. Hotelling's T^2 control chart is a multivariate control chart commonly used to monitor the mean process. The performance of the T^2 chart in monitoring mean shifts can be increased if a robust estimator is utilized. Based on previous research, T^2 based on the Fast-MCD estimator has good performance in monitoring low to medium outlier contaminated data. Therefore, the MRCD estimator can be used to detect intrusion. On the other hand, this research focuses on developing a bootstrap-based robust Hotelling's T^2 charts with Fast-MCD and MRCD estimators for evaluating performance in detecting intrusion on intrusion detection datasets. Based application of UNSW-NB15, the proposed chart has better performance than the conventional T^2 and Fast-MCD-based T^2 despite the longer execution time.

Keywords: bootstrap; intrusion detection; multivariate control chart; MRCD; Hotelling's T^2

1. Introduction

Intrusion detection is a process of monitoring events occurring within a computer system or network, followed by analyzing the monitoring data to identify indications of intrusion attempts. Intrusion refers to attempts to gain unauthorized access to a computer system or network, potentially threatening the availability, integrity, and confidentiality of a computer network system. The system used to perform intrusion detection is known as an Intrusion Detection System (IDS) [1]. Intrusion detection is generally carried out by matching network traffic patterns with known attack patterns or by identifying abnormal network traffic patterns [2]. In general, anomaly-based network intrusion detection systems are categorized into three categories: knowledge-based systems, computational approaches, and statistical approaches [3]. One statistical methodological approach used in intrusion detection is Statistical Process Control (SPC), widely applied across various sectors including industries and services. Besides detecting changes in manufacturing and service processes, SPC can also be applied in IDS. Research has explored the application of SPC in the context of intrusion detection [4].

Statistical Process Control (SPC) has played a major role in product quality control since Shewhart [5] introduced the control chart techniques by applying statistical methods to monitor the industrial processes. One of the multivariate control charts which is commonly used to monitor the process mean is Hotelling's T^2 control chart [6], which can be used to monitor either individual or subgroup observations. In SPC concepts, an outlier can be defined as an observation that significantly deviates from other observations, which indicates that the observation is observed by a different process [7]. The Hotelling's T^2 chart is not suitable to detect the presence of multiple outlier [8], due

 doi:10.20944/preprints202312.2062.v1

to the masking and swamping effect [9], especially for highly outlier contaminated data. The statistic of T^2, which is based on the classical estimator, is easily affected and decreased by the presence of outliers [10,11]. Moreover, the performance of control charts will decrease if the variables monitored increase [12].

To overcome those problems, several methods have been proposed to minimize the effects of outliers by changing the classical estimator with a robust estimator, especially for the covariance matrix estimator. The performance of the T^2 control chart in monitoring mean shifts will increase if a robust estimator is utilized [13]. Many robust methods have been adopted to develop a T^2 control chart to minimize the effect of outliers. These methods such as Minimum Volume Ellipsoid (MVE) [14], Trimming Method [15,16], Minimum Vector Variance [17,18], Successive Difference Covariance Matrix (SDCM) [10,19], Minimum Covariance Determinant (MCD) [15,20], Reweighted minimum covariance determinant (RMCD) [21], and Fast Minimum Covariance Determinant (Fast-MCD), whose good performance on monitoring small to medium outlier contaminated data with 30% breakdown point [22]. The latest development of robust estimators is the Minimum Regularized Covariance Determinant (MRCD) method [23], which uses the concept of data-driven algorithm and regularization to avoid overfitting problems. The MRCD estimator can be used to detect outliers in high-dimensional data. Besides the robust estimator, the Hotelling' T^2 chart can also be developed using a non-parametric approach as a control limit, namely the bootstrap resampling method [24].

This research focuses on developing bootstrap-based robust T^2 control charts with MRCD estimators for detecting intrusion. This method will be applied to the UNSW-NB15 dataset. The rest of this paper is organized as follows: Section 2 presents the related work. In Section 3, the explanation of the proposed chart construction is presented. Section 4 provides the methodology and procedure of the proposed chart. Section 5 shows the application results of the proposed chart for the IDS dataset. Finally, Section 6 is allocated for the conclusion and future research.

2. Related Works

The SPC method commonly used in intrusion detection is a multivariate control chart. Ye et al. [25] initiated the use of Markov Chain techniques, T^2 Hotelling, and chi-square multivariate tests for intrusion detection. Then Ye et al. [26] proposed a technique based on Hotelling's T^2 that can detect both counter relationships and mean-shift anomalies. Qu, Hariri, and Yousif [27] use the T^2 Hotelling diagram to detect intrusions on a network called real-time Multivariate Analysis for the Network Attack detection algorithm (MANA) by updating control limits at certain time intervals. Zhang, Zhu, and Jin [28] developed a Support Vector Clustering (SVC) based control diagram with performance results similar to the T^2 diagram for detecting anomalies in computer networks. Tavallaee et al. [29] apply Covariance Matrix Sign (CMS) to detect Denial of Service (DoS) attacks. Sivasamy and Sundan [30] compared the performance of the T^2 Hotelling control chart with the SVM and TANN methods and found that Hotelling's T^2 accuracy level was high for all types of attack classes.

In addition to Hotelling's T^2, Rastogi et al. *[31]* stated that in theory MEWMA and MCUSUM can be used in intrusion detection, however, intrusion detection data involves many quality characteristics so MEWMA and MCUSUM are not suitable for use. Camacho et al [32] use PCA based on Multivariate Statistical Process Control (MSPC) to monitor intrusions. Ahsan et al. [33] use PCA-based Hotelling's T^2 which produces more efficient computational time. The use of non-parametric control limits improves performance on the T^2 control diagram with a Successful Difference Covariance Matrix (SDCM) in the form of Kernel Density Estimation [34] and Bootstrap Resampling *[35]*. Then Ahsan et al. [36] developed robust Hotelling's T^2 based on Fast-MCD which shows better performance in detecting outliers in intrusion detection systems.

3. Proposed Chart

In this section, the procedures of the proposed chart are explained. The Minimum Regularized Covariance Determinant (MRCD) estimator is employed to enhance the robustness of the mean vector and covariance matrix on the chart's performance for intrusion detection.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2023 doi:10.20944/preprints202312.2062.v1

3.1. Multivariate Hotelling's T^2 Control Chart

In this section, a short brief of the conventional Hotelling's T^2 control chart is shown. Hotelling's T^2 is a multivariate control chart, a generalization of the *t*-student distribution, that can be used to monitor the process mean. Let $\boldsymbol{x}_i$ where $i = 1, 2, \dots, n$ are identic and independently random vectors which follow the multivariate normal distribution $\boldsymbol{x}_i \sim N_p(\mu, \boldsymbol{\Sigma})$. The data structure can be written as $\boldsymbol{X} = [\boldsymbol{x}_1^T \quad \boldsymbol{x}_2^T \quad \cdots \quad \boldsymbol{x}_n^T]$ with mean vector $\bar{x} = \frac{1}{n}\sum \boldsymbol{x}_i$ and covariance matrix $\boldsymbol{S} = \frac{1}{n-1}\sum (x_i - \bar{x})(x_i - \bar{x})^T$. The T_i^2 statistics can be calculated as follows [6]:

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

Conventional Hotelling's T^2 chart follows the assumption of multivariate normal distribution [37], so the control limit can be generated by following *F*-distribution with the equation:

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

where n is the total number of observations p is the variables quantity with α is the false alarm rate. The monitoring process is said in-control if T^2 statistics are not greater than the control limit.

3.2. Bootstrap Control Limit-based T^2 Chart

In some cases, a random variable might not follow any certain distributions. To overcome this problem, the bootstrap method can be applied to estimate the parameter of unknown distribution [38,39]. Despite initially proposed on classic T^2 statistics [24], this control limit also can be adopted on the proposed chart by putting robust statistics on the first step. The algorithm of bootstrap control limit calculation (see Figure 1 for illustration) is presented as follows:

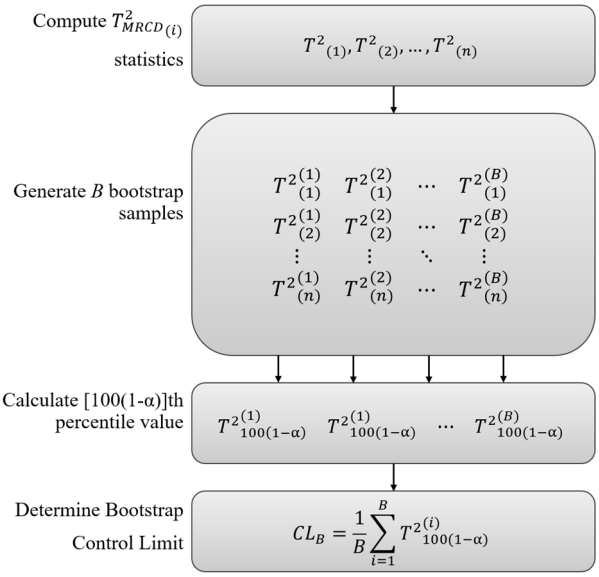


Figure 1. Bootstrap Control Limit Algorithm.

Algorithm of Bootstrap Control Limit

Step 1. Compute the statistic T^2 with n observations

Step 2. Generate B times bootstrap samples from statistic T^2 for n observations with replacement (e.g., B=1,000).

Step 3. Calculate 100(1-α).th percentile for each bootstrap resample for statistic $T^{2(l)};\ l = 1, 2, \dots, n$

Step 4. Determine the bootstrap control limit by averaging each replication using

$$CL_B = \frac{1}{B}\sum_{i=1}^{B} T^{2(l)}_{100(1-\alpha)}$$

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2023 doi:10.20944/preprints202312.2062.v1

3.2. MRCD Algorithm

The minimum Covariance Determinant (MCD) based method is the most widely used robust estimator of multivariate location and scatters [40]. This method is designed to determine $\boldsymbol{H}_{MCD}$ defined as subset with the smallest sample covariance determinant. The MCD estimates for the mean vector and covariance matrix correspond to the mean vector and covariance matrix of $\boldsymbol{H}_{MCD}$. Define h as the subset size where $\frac{n}{2} \leq h < n$, and $h \geq p$ must be fulfilled, otherwie the MCD covariance matrix will be singular. The MCD algorithm calculates every subset possible as many as $\binom{n}{h}$ possible combinations in order to get $\boldsymbol{H}_{MCD}$. So, this method is time-consuming and not suitable for estimating large datasets.

Minimum Regularized Covariance Determinant (MRCD) estimators are proposed [23] as the extension of MCD. The MRCD is a robust estimator that uses various combinations of target matrix and regularization weight determined through data-driven procedures. The application of the MRCD estimator has good robustness and can be used to deal with outliers in high-dimensional data.

Let $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_i, \dots, \boldsymbol{x}_n\}'$ where $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \dots, x_{ip}\}'$ from a p-variate observations. First, the data need to be standardized using the Qn estimator [41] and put these values in a diagonal matrix $\boldsymbol{D}_X$. The median of each variable also needs to be computed and put in a location vector $\boldsymbol{v}_X$. The stardardized obesvations are then stated on $\boldsymbol{U}$ that constructed under a set of $\boldsymbol{u}_i$ as follows:

$$\boldsymbol{u}_i = \boldsymbol{D}_X^{-1}(\boldsymbol{x}_i - \boldsymbol{v}_X) \tag{3}$$

The next step is defining the Target Matrix (**T**) and scalar regularization parameter(ρ). **T** is a $p \times p$ diagonal matrix that consists of estimated univariate scales, while ρ is a weighted parameter that can be obtained by the data-driven approach which satisfies $0 \leq \rho \leq 1$. Then define $h \times p$ covariance matrix $\boldsymbol{K}(H)$ of h-subset of H on the standardized data $\boldsymbol{U}$ as follows:

$$\boldsymbol{K}(H) = \rho\, \boldsymbol{T} + (1-\rho) c_\gamma \boldsymbol{S}_{\boldsymbol{U}}(H) \tag{4}$$

where $\boldsymbol{S}_{\boldsymbol{U}}(H)$ is the covariance matrix of the h-subset for $\boldsymbol{U}$ and c_γ is the consistency factor [42]

Mathematical operation (7) can be done with a spectral decomposition $\mathbf{T} = \mathbf{Q}\boldsymbol{\lambda}\mathbf{Q}'$ where $\boldsymbol{\lambda}$ is the diagonal matrix containing the eigenvalues of **T**, while **Q** is the orthogonal matrix containing the corresponding eigenvectors. The previous equation can be rewritten as follows

$$\boldsymbol{K}(H) = \boldsymbol{Q}\boldsymbol{\lambda}^{1/2}[\rho\, \boldsymbol{I} + (1-\rho) c_\alpha \boldsymbol{S}_{\boldsymbol{W}}(H)]\boldsymbol{\lambda}^{1/2}\boldsymbol{Q}' \tag{5}$$

where **W**, containing the transformation of standardized observations $\boldsymbol{w}_i = \boldsymbol{\lambda}^{\mathbf{1/2}}\mathbf{Q}'\boldsymbol{u}_i$. Consequently, it follows $\boldsymbol{S}_{\boldsymbol{W}}(H) = \boldsymbol{\lambda}^{\mathbf{-1/2}}\mathbf{Q}'\boldsymbol{S}_{\boldsymbol{U}}(H)\mathbf{Q}\boldsymbol{\lambda}^{\mathbf{-1/2}}$.

The subset MRCD $\boldsymbol{H}_{MCD}$ obtained by minimizing the determinant of the regularized covariance matrix **K***(H)* as:

$$\begin{aligned} \boldsymbol{H}_{MRCD} &= \underset{H \in \Omega}{\operatorname{argmin}}\big(\det(\boldsymbol{K}(H))\big) \\ \boldsymbol{H}_{MRCD} &= \underset{H \in \Omega}{\operatorname{argmin}}\big(\det\big(\rho\, \mathbf{I} + (1-\rho) c_\alpha \boldsymbol{S}_{\boldsymbol{W}}(H)\big)\big) \end{aligned} \tag{6}$$

Once the $\boldsymbol{H}_{MRCD}$ is determined, then the location and scatters of the MRCD estimator can be defined as.

$$\bar{\boldsymbol{x}}_{\grave{MRCD}} = \boldsymbol{v}_X + \boldsymbol{D}_{\boldsymbol{X}}\bar{\boldsymbol{H}}_{MRCD} \tag{7}$$

$$\boldsymbol{S}_{MRCD} = \boldsymbol{D}_{\boldsymbol{X}}\mathbf{Q}\boldsymbol{\lambda}^{1/2}[\rho\, \mathbf{I} + (1-\rho) c_\alpha \boldsymbol{S}_{\boldsymbol{W}}(\boldsymbol{H}_{MRCD})]\boldsymbol{\lambda}^{1/2}\mathbf{Q}'\boldsymbol{D}_{\boldsymbol{X}} \tag{8}$$

3.3. MRCD-Based T^2 Chart

In order to develop the robust Hotelling's T^2 control chart, this study changes the classic estimators of mean vectors $\bar{\boldsymbol{x}}$ and covariance matrix $\boldsymbol{S}$ from equation (1) with the estimated value of mean vector and covariance matrix from robust estimators. Robust T^2 statistic based on MRCD was constructed as follows

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2023 doi:10.20944/preprints202312.2062.v1

$$T^2_{MRCD;i} = (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{MRCD})' \mathbf{S}^{-1}_{MRCD} (\boldsymbol{x}_i - \bar{\boldsymbol{x}}_{MRCD}) \quad (9)$$

Due to the unknown distribution of the proposed chart, its control limit for both charts are estimated using the bootstrap resampling method to develop an adaptive control chart. The detailed procedure for the control limit is presented in the previous subsection.

4. Methodology

In developing the proposed robust T^2 chart based on the MRCD estimator, there are two phases required to be undertaken. Phase I is building a normal profile from the in-control or the normal profile, while Phase II is detecting the intrusion using the calculated statistics and control limit from Phase II. Phase I needs to calculate the mean vector, covariance matrix, and bootstrap control limit. The procedure of Phase I is shown as following these steps:

Phase I: Building Normal Profile

Step 1. Form the in-control or normal data matrix $\boldsymbol{X}_{normal}$

Step 2. Calculate $\bar{\boldsymbol{x}}_{MRCD}$ and $\boldsymbol{S}_{MRCD}$, which are the robust estimated values of normal data $\boldsymbol{X}_{normal}$ using the MRCD algorithm in equation (11) and (12)

Step 3. Calculate $T^2_{MRCD;i}$ using equation (14) from normal data $\boldsymbol{X}_{normal}$

Step 4. Determine α and compute the bootstrap control limit $CL_{B;MRCD}$

Then the estimated normal profile and control limit from Phase I are utilized in the detection process in Phase II. The procedure of Phase II is shown as follows:

Phase II: Detection

Step 1. Form the new data matrix $\boldsymbol{X}_{test}$

Step 2. Calculate $T^2_{MRCD;i}$ from new data $\boldsymbol{X}_{normal}$ as follows:
$T^2_{MRCD;i} = (\boldsymbol{x}_{test;i} - \bar{\boldsymbol{x}}_{MRCD})^T \mathbf{S}^{-1}_{MRCD} (\boldsymbol{x}_{test;i} - \bar{\boldsymbol{x}}_{MRCD})$
where $\bar{\boldsymbol{x}}_{MRCD}$ and $\boldsymbol{S}_{MRCD}$ are taken from Phase I

Step 3. Detect if $T^2_{MRCD;i} > CL_{B;MRCD}$ then the observation is labeled as an intrusion and if $T^2_{MRCD;i} \leq CL_{B;MRCD}$ then the observation is labeled as normal

Moreover, the performance of the proposed can be assessed and evaluated by the confusion matrix table shown in Table 1. The classification goodness method could be measured by the degree of goodness and degree of error. The goodness in intrusion detection can be divided into two types:

Table 1. Confusion Matrix Table.

Actual	Detection	
	Intrusion	**Normal**
Intrusion	True Positives (TP)	False Negatives (FN)
Intrusion	False Positives (FP)	True Negatives (TN)

a. True Positives (TP) are intrusion records that are successfully detected as intrusion.
b. True Negatives (TN) are normal records that are correctly stated as normal.

The errors in detecting intrusion also can be divided into two types:

a. False Positives (FP) are normal records that are incorrectly detected as intrusions.
b. False Negatives (FN) is an intrusion records that unsuccessfully detected as normal records.

FP leads to a false alarm while FN results in an undetected intrusion on the chart. Those types of error can be used to calculate the degree of error namely FP Rate and FN Rate [43]. While the level of goodness can be measured using the Area Under Curves (AUC) as follows [44]:

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2023 doi:10.20944/preprints202312.2062.v1

$$AUC = \frac{1}{2}\left(\frac{TP}{TP + FN} + \frac{TN}{TN + FP}\right) \quad (10)$$

$$FP\ Rate = \frac{FP}{TN + FP} \quad (11)$$

$$FN\ Rate = \frac{FN}{TP + FN} \quad (12)$$

5. Results and Discussions

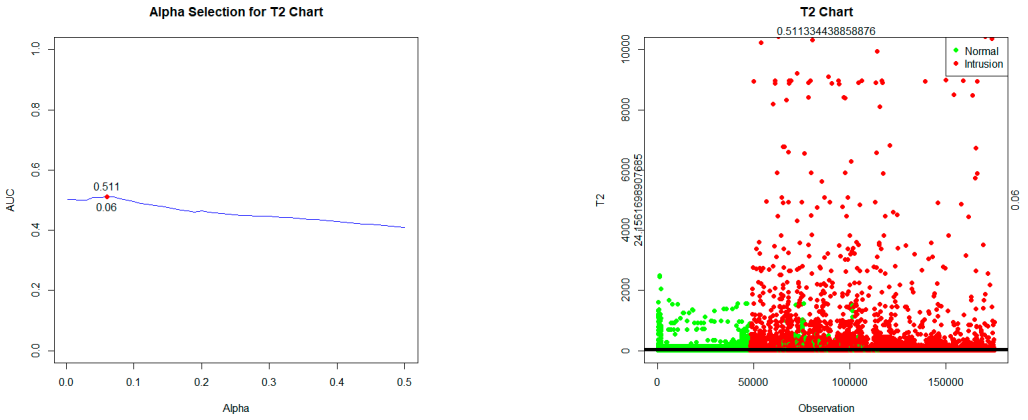
The UNSW-NB15 dataset was built using the IXIA PerfectStorm tool at the Australian Centre for Cyber Security (ACCS) by generating a combination of normal activities and realistic, modern artificial attacks for research purposes related to Network Intrusion Detection Systems (NIDS) [45]. Compared to other NIDS datasets, UNSW-NB15 excels in complexity, referring to patterns of modern network traffic attacks, making it suitable for evaluating intrusion detection systems [46]. The training set of UNSW-NB15 consists of 175,341 records with 38 metric features and record labels which are normal labels and several types of intrusion labels that are presented in Table 2.

Table 2. Characteristics of UNSW-NB15 Dataset.

Label	Number of Records	Percentage
Normal	56.000	31,94
Intrusion	119.341	68,06
-Analysis	2.000	1,14
-Backdoor	1.746	1,00
-DoS	12.264	6,99
-Exploits	33.393	19,04
-Fuzzers	18.184	10,37
-Generic	40.000	22,81
-Reconnaissance	10.491	5,98
-Shellcode	1.133	0,65
-Worms	130	0,07
Total	175.341	100,00

The data application is conducted through three methods: conventional Hotelling's T^2, robust T^2 based on Fast-MCD, and the proposed diagram, which is the robust T^2 based on MRCD. The construction of the control chart is divided into two phases: Phase I for establishing control limits and Phase II for the detection process and calculating the performance of the control chart. In the conventional Hotelling's T^2 control chart, the T^2 statistic is calculated using equation (2.29), with control limits determined based on the significance level using the criteria of the highest AUC value, which is α=6%, as depicted in Figure 2(a). After computing the statistics and establishing control limits, the control chart can be visualized, as shown in Figure 2(b).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2023 doi:10.20944/preprints202312.2062.v1



(a) (b)

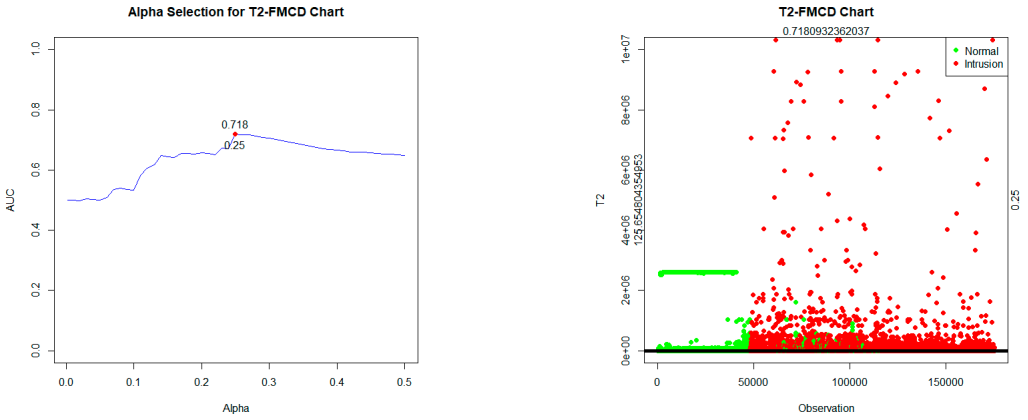
Figure 2. T^2 Chart of (a) Control Limit Selection and (b) Statistic Plot.

Based on Figure 2, the statistical plot depicts two types of data labels: green for normal data and red for intrusion data. These statistics will be tested against the control limits. If the value of the statistic $T^2 > CL_B$, the observation is detected as an intrusion. While if the statistic $T^2 \leq CL_B$, the observation is detected as normal. Based on the labels and the detection outcomes obtained, a confusion matrix table can be formed in Table 3.

Table 3. Confusion Matrix and Performance Evaluation for Conventional Hotelling's T^2 Chart

Actual	Detection		Accuracy	AUC	FP Rate	FN Rate
	Intrusion	Normal				
Intrusion	9,853	109,488	0.376	0.511	0.060	0.917
Normal	3,354	52,646				

Next, in the construction of a control chart for Robust T^2 based on Fast-MCD, the T^2 statistic is calculated, and the control limits are determined based on the significance level using the criteria of the highest AUC value, which is α=25%, as depicted in Figure 3(a). After computing the statistics and establishing the control limits, the control chart can be visualized, as seen in Figure 3(b).



(a) (b)

Figure 3. Fast-MCD-based T^2 Chart of (a) Control Limit Selection and (b) Statistic Plot.

Based on the figure, the statistical plot depicts two types of data labels: green for normal data and red for intrusion data. These statistics will be tested against the control limits. If the value of the statistic $T^2_{FMCD} > CL_B$, the observation is detected as an intrusion. While if the statistic $T^2_{FMCD} \leq CL_B$,

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 27 December 2023 doi:10.20944/preprints202312.2062.v1

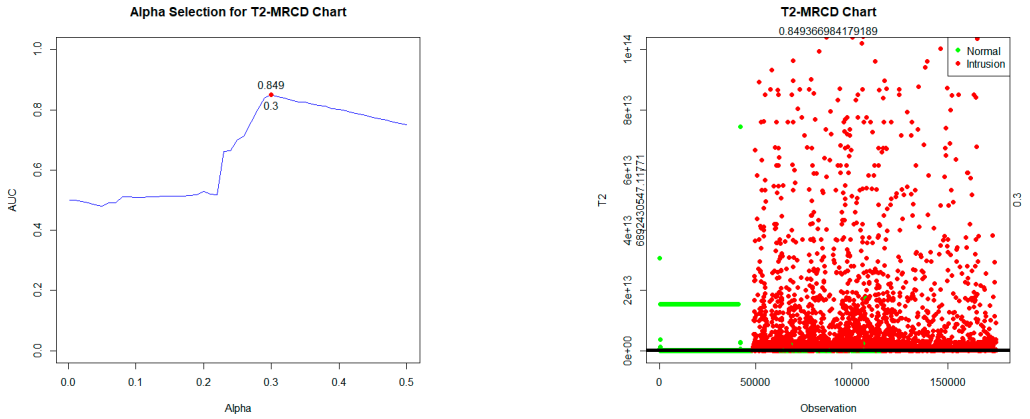
the observation is detected as normal. Based on the labels and the detection outcomes obtained, a confusion matrix table can be formed and evaluated as Table 4.

Table 4. Confusion Matrix and Performance Evaluation for Fast-MCD-based T^2 Chart

Actual	Detection		Accuracy	AUC	FP Rate	FN Rate
	Intrusion	Normal				
Intrusion	81,871	37,470	0.711	0.718	0.250	0.314
Normal	13,991	42,009				

Based on Table 4, it can be known that the performance of Robust T^2 based on Fast-MCD on the UNSW-NB15 data is quite good, with an AUC value of 0.718. Additionally, with an FP Rate of 0.25, there's a relatively low FN rate of 0.314.

For constructing the proposed chart of Robust T^2 based on MRCD, the T^2 statistic is calculated using, and the control limits are determined based on the significance level using the criteria of the highest AUC value, which is α=30%, as depicted in Figure 4(a). After computing the statistics and establishing control limits, the control chart can be visualized, as shown in Figure 4(b).



(a) (b)

Figure 4. MRCD-based T^2 Chart of (a) Control Limit Selection and (b) Statistic Plot.

Based on Figure 4, the statistical plot depicts two types of data labels: green for normal data and red for intrusion data. These statistics will be tested against the control limits. If the value of the statistic $T^2_{MRCD} > CL_B$, the observation is detected as an intrusion. While if the statistic $T^2_{MRCD} \leq CL_B$, the observation is detected as normal. Based on the labels and the detection outcomes obtained, a confusion matrix table can be formed and evaluated in Table 5.

Based on Table 5, it's apparent that the performance of Robust T^2 based on MRCD on the UNSW-NB15 data is excellent, with an AUC value of 0.849. Additionally, with an FP Rate of 0.298, there's an exceptionally low FN rate of only 0.004.

Table 5. Confusion Matrix and Performance Evaluation for MRCD-based T^2 Chart

Actual	Detection		Accuracy	AUC	FP Rate	FN Rate
	Intrusion	Normal				
Intrusion	118,915	426	0.902	0.849	0.298	0.004
Normal	16,671	39,329				

 doi:10.20944/preprints202312.2062.v1

After applying the UNSW-NB15 data using these three methods, the performance of each chart can be compared and evaluated based on several goodness and error criteria, as presented in Table 6.

Table 6. Confusion Matrix and Performance Evaluation for MRCD-based T^2 Chart

Control Chart	Accuracy	AUC	FP Rate	FN Rate	Execution Time (s)
Conventional T^2	0.376	0.511	0.060	0.917	286
T^2 Fast-MCD	0.711	0.718	0.250	0.314	1,470
T^2 MRCD	0.902	0.849	0.298	0.004	8,108

Table 6 displays the Accuracy, AUC, FP Rate, FN Rate, and execution time of the three methods used in this study. The conventional T^2 method, with its straightforward steps, took only 286 seconds. The Fast-MCD-based T^2 method, known for its efficiency, required 1,470 seconds. Meanwhile, the MRCD-based T^2, featuring a complex algorithm, took a longer time of 8,108 seconds.

The duration of execution time correlates with the quality of the chart's performance in detecting intrusions. Based on the AUC values, the conventional T^2 Hotelling chart showed poor performance in intrusion detection, achieving an AUC of only 0.511. Both robust T^2 charts demonstrated better performance than the conventional T^2. The Fast-MCD-based T^2 had a relatively good AUC value of 0.718. On the other hand, the proposed MRCD-based T^2 had the best performance with the highest AUC value of 0.849 and an exceptionally low FN Rate of 0.004, indicating a very low chance of undetected intrusions.

6. Conclusion and Future Research

The application of the UNSW-NB15 data revealed that the MRCD-based T^2 Hotelling exhibited better performance in detecting intrusion, with a 0.902 Accuracy and 0.848 AUC value. This proposed chart successfully outperformed the conventional T^2 Hotelling and the Fast-MCD-based T^2 Hotelling, which had AUC values of 0.511 and 0.718, respectively despite the longer execution time. For further research, this proposed chart still can be modified by applying another non-parametric approach as the control limit. Applying an MRCD estimator for monitoring processes for variance shifts or simultaneous shifts in mean and variance also can be constructed. The latest robust estimator, Cellwise MCD [47], known for its efficiency can be considered to be implemented to overcome the MRCD's problem in terms of long duration of execution time.

Author Contributions: I.K.P.: writing original draft and data analysis. M.A.: Conceptual methodology, Supervising and validating the results. M.M.: Performed analysis and data visualization. M.H.L.: Validating the results.

Acknowledgments: The authors gratefully acknowledge financial support from the Institut Teknologi Sepuluh Nopember for this work, under the project scheme of the Publication Writing and IPR Incentive Program (PPHKI) 2022

Conflicts of Interest: The authors declare no conflict of interest.

References

1. R. Bace and P. Mell, "NIST special publication on intrusion detection systems.," *Nist Special Publication*, 2001.
2. W. Lee and S. J. Stolfo, "A framework for constructing features and models for intrusion detection systems," *ACM Transactions on Information and System Security (Vol. 3)*, 2000.
3. G. Wu and Y. Huang, "Design of a New Intrusion Detection System Based on Database," *2009 International Conference on Signal Processing Systems*, pp. 814-817, 2009.
4. Y. Park, A Statistical Process Control Approach for Network Intrusion Detection, Georgia Insitute of Technology, 2005.

5. W. A. Shewhart, "Some applications of statistical methods to the analysis of physical and engineering data, 3(1)," *Bell System Technical Journal,* p. 43–87, 1924.
6. H. Hotteling, Multivariate Quality Control. In Techiques of Statistical Analysis, New York: McGraw-Hill, 1947.
7. D. M. Hawkins, Identification of outliers. Vol. 11, Berlin, Germany: Springer, 1980.
8. S. Chenouri, S. H. Steiner and A. M. Variyath, "A multivariate robust control chart for individual observations," *Journal of Quality Technology 41 (3),* p. 259–271, 2009.
9. P. J. Rousseeuw and B. C. Van Zomeren, "Unmasking multivariate outliers and leverage point," *Journal of the American Statistical Association 85 (411),* p. 633–9, 1990.
10. J. H. Sullivan and W. H. Woodall, "A comparison of multivariate control charts for individual observations," *Journal of Quality Technology 28 (4),* p. 398–408, 1996.
11. P. J. Rousseeuw and A. M. Leroy, Robust regression and outlier detection. Vol. 589, John Wiley & Sons, 2005.
12. M. Ahsan and H. Khusna, "Evaluasi Diagram Kontrol Multivariat berbasis Independen Principal Component Analysis (PCA)," *INFERENSI,* vol. 1, no. 2, pp. 89-92, 2018.
13. G. Williems, G. Pison, P. J. Rousseeuw and S. Van Aelst, "A Robust Hotelling Test," *Metrika,* pp. 125-138, 2002.
14. N. J. Vargas, "Robust estimation in multivariate control charts for individual observations," *Journal of Quality Technology,* vol. 35, no. 4, p. 367–376, 2003.
15. J. L. Alfaro and J. F. Ortega, "A comparison of robust alternatives to hotelling's T2 control chart," *Journal of Applied Statistics 36 (12),* p. 1385–1396, 2009.
16. M. O. A. Abu-Shawiesh, B. M. Golam Kibria and F. George, "A robust bivariate control chart alternative to the hotelling's T2 control chart," *Quality and Reliability Engineering International 30 (1),* p. 25–35, 2014.
17. S. S. S. Yahaya, H. Ali and Z. Omar, "An alternative hotelling T 2 control chart based on minimum vector variance (MVV)," *Modern Applied Science,* vol. 5, no. 4, p. 132, 2011.
18. D. E. Herwindiati, M. A. Djauhari and M. Mashuri, "Robust multivariate outlier labeling," *Communications in Statistics Simulation and Computation,* vol. 36, no. 6, p. 1287–94, 2007.
19. J. D. Williams, W. H. Woodall, J. B. Birch and J. O. E. H. Sullivan, "On the distribution of hotelling's T2 statistic based on the successive differences covariance matrix estimator," *Journal of Quality Technology,* vol. 38, no. 3, p. 217–229, 2006.
20. W. A. Jensen, J. B. Birch and W. H. Woodall, "High breakdown estimation methods for phase I multivariate control charts," *Quality and Reliability Engineering International,* vol. 23, no. 5, p. 615–629, 2007.
21. A. N. F. Utami and Suwanda, "Penggunaan Estimator Robust Reweighted Minimum Covariance Determinant pada Diagram Kontrol T2 Hotelling untuk Monitoring Penyebaran Covid-19 di Korea Selatan," *Jurnal Riset Statistika,* pp. 63-72, 2021.
22. M. Ahsan, M. Mashuri, M. H. Lee, H. Kuswanto and D. D. Prastyo, "Robust Adaptive Multivariate Hotelling T2 Control Chart Based on Kernel Density Estimation for Intrusion Detection System," *Expert Systems with Applications,* vol. 145, no. 113105, 2020.
23. K. Boudt, . P. J. Rousseeuw, S. Vanduffel and T. Verdonck, "The minimum regularized covariance determinant estimator.," *Statistics and Computing,* vol. 30, no. 1, p. 113–28, 2020.
24. P. Phaladiganon, S. B. Seoung, V. C. P. Chen, J. G. Baek and S. K. Pa, "Bootstrap-Based T2 Multivariate Control Charts," *Communications in Statistics - Simulation and Computation,* vol. 40, no. 5, pp. 645-662, 2011.
25. N. Ye, X. Li, Q. Chen, S. M. Emran and M. Xu, "Probabilistic techniques for intrusion detection based on computer audit data," *IEEE Trans. Syst. Man, Cybern. Part ASystems Humans,* vol. 31, p. 266–274, 2001.
26. N. Ye, S. M. Emran, Q. Chen and S. Vilbert, "Multivariate statistical analysis of audit trails for host-based intrusion detection," *IEEE Transactions on Computers,* vol. 51, no. 7, p. 810–820, 2002.
27. G. Qu, S. Hariri and M. Yousif, "Multivariate Statistical Analysis for Network Attacks Detection," *The 3rd ACS/IEEE International Conference onComputer Systems and Applications,* p. 9–14, 2005.
28. Z. Zhang, X. Zhu and J. Jin, "SVC-Based Multivariate Control Charts for Automatic Anomaly Detection in Computer Networks," *IEEE,* 2007.
29. M. Tavallaee, W. Lu, S. A. Iqbal and A. Ghorbani, "A Novel Covariance Matrix based Approach for Detecting Network Anomalies," *In Sixth Annual Conference on Communication Networks and Services Research,* 2008.

30. A. Sivasamy and B. Sundan, "A Dynamic Intrusion Detection System Based on Multivariate Hotelling's T2 Statistics Approach for Network Environments," *The Scientific World Journal,* p. 1–9, 2015.
31. R. Rastogi, Z. Khan and M. H. Khan, "Network Anomalies Detection Using Statistical Technique : A Chi-Square approach.," *International Journal of Computer Science Issues,* vol. 9, no. 2, p. 515–522, 2012.
32. J. Camacho, A. Pérez-Villegas, P. García-Teodoro and G. Maciá-Fernández, "PCA-based multivariate statistical network monitoring for anomaly detection," *Computers & Security,* vol. 59, pp. 118-137, 2016.
33. M. Ahsan, M. Mashuri, H. Kuswanto and D. Prastyo, "Intrusion detection system using multivariate control chart Hotelling's T2 based on PCA," *International Journal on Advanced Science, Engineering and Information Technology,* vol. 8, no. 5, pp. 1905-1911, 2018.
34. M. Ahsan, M. Mashuri, H. Kuswanto, D. D. Prastyo and H. Khusna, "T2 control chart based on successive difference covariance matrix for intrusion detection system," *Journal of Physics: Conference Series,* vol. 1028, 2018.
35. M. Ahsan, M. Mashuri and H. Khusna, "INTRUSION DETECTION SYSTEM USING BOOTSTRAP RESAMPLING APPROACH OF T^2 CONTROL CHART BASED ON SUCCESSIVE DIFFERENCE COVARIANCE MATRIX," *Journal of Theoretical & Applied Information Technology,* vol. 96, no. 8, 2018.
36. M. Ahsan, M. Mashuri, M. H. Lee and H. Kuswanto, "Robust Adaptive Multivariate Hotelling T2 Control Chart Based on Kernel Density Estimation for Intrusion Detection System," *Expert Systems with Applications,* vol. 145, no. 113105, 2020.
37. D. C. Montgomery, Introduction to Statistical Quality Control, 7th edition., New York: John Wiley & Sons, Inc,, 2013.
38. B. Efron, "Bootstrap Methods: Another Look at the Jacknife," *Ann. Stat.,* vol. 7, pp. 1-26, 1979.
39. B. Efron and R. Tibshirani, "An introduction to the bootstrap," *CRC Press,* 1994.
40. P. J. Rousseeuw, "Multivariate estimation with high breakdown point," *Mathematical Statistics and Applications,* vol. 8, no. 37, p. 283–297, 1985.
41. P. Rousseeuw and C. Croux, "Alternatives to the median absolute deviation," *J. Am. Stat. Assoc.,* vol. 88, no. 424, p. 1273–1283, 1993.
42. C. Croux and G. Haesbroeck, "Influence function and efficiency of the minimum covariance determinant scatter matrix estimator.," *Journal of Multivariate Analysis,* vol. 71, no. 2, pp. 161-190, 1999.
43. J. Han, M. Kamber and J. Pei, Data Mining Concepts and Techniques, USA: Morgan Kaufmann, 2012.
44. M. Bekkar, H. Djemaa and T. Alitouch, "Evaluation Measure for Models Assesment Over Imbalanced Data Sets," *Journal of Information Engineering and Aplications,* pp. 27-38, 2013.
45. N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection," *2015 Military Communications and Information Systems Conference. Canberra, Australia: MilCIS 2015-IEEE Stream,* pp. 1-6, 2015.
46. N. Moustafa and J. Slay, "The evaluation of Network Anomaly Detection Systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," *Information Security Journal: A Global Perspective,* vol. 25, no. 1-3, pp. 18-31, 2016.
47. J. Raymaekers and P. J. Rousseeuw, "The Cellwise Minimum Covariance Determinant Estimator," *Journal of the American Statistical Association,* 2023.