# CHAPTER 1
# INTRODUCTION

Background of this research is discussed on section 1.1. In section 1.2, we talk about the concept and overall aspects that will affect the research. Section 1.3 explain about problem that this research trying to solve, while in section 1.4 describe the hypothesis on some aspects on this research. Meanwhile, lists of assumptions that used in this research are described in 1.5, while in Section 1.6 we discuss about scope and delimitation of this research. Contribution of this research is stated on section 1.7.

## 1.1 Rationale

The increasing number of attacks against computer networks, make security becomes an important issue in this system. Every year, security threats that encountered the network becomes more sophisticated, organized and difficult to detect. At the same time, the consequences of blocking failure of attacks are increase. This is a violation of computers security policy, i.e. *Confidentiality*, *Integrity* and *Availibility* (CIA) [2]. In the real world, according to survey "Threat Report" from McAfee Labs., in 2016 recorded attacks appears the most are Browser (33%), Brute Force (23%), Denial-of-Service (22%), SSL (7%), DNS (4%), Scan (3%), Backdoor (3%), Others (5%). This is indicate that the issue of computer networks is still a research topic that still requires optimal solution.

To resolve this problem, a better security system against these attacks should be constructed, both in terms of speed, accuracy and reliability of detection. The purpose of computer security is to support a user through malicious or detect a intrusion in system information. One of security system method is Intrusion Detection System (IDS). IDS is security system that monitors its system from inbound and outbound attacks. IDS will receive copies of the package which addressed to a host to further examination what kind of package is. If it found dangerous package, then the IDS would alert the system administrator. According to [2], IDS can inspect inbound and outbound traffic in a system or network, conduct an analysis and looking evidence of attempted intrusion. As well as anti-virus, firewall and access control, IDS can be used to strengthen the security

of the system [3]. IDS can be able to recognize deviations or abnormal usage on the network assets, by collecting information and data from system logs to be analyzed so that it can be determined steps against an infiltration that were found. The existing IDS problem is difficult to recognize normal user or intruders [4]. Some problems which faced by IDS is a high false positive and low detection accuracy [5]. False positive rate (FPR) is the proportion of normal traffic detected falsely as attack. False positive rate problems may bother an administrator. So that, the administrator tasks become more complex. They need to arrange which one is the real attack of the many reports of false positives that arise. Accuracy refers to the proportion of true detected with total data [1].

The detection method on IDS is divided into three, i.e. anomaly-based, signature-based and state protocol analysis (SPA). Signature-based is the method based on the known attacks. The detection is done by comparison with a database that contains information about the type of intrusion. Therefore, the Signature-Based also called knowledge-based or misuse-based detection. This method is effective to detect these types of attacks are recorded in the database, but can not survive against zero-day attacks [6]. Anomaly-based is a method based on deviation of normal behavior patterns. This method is effective against this type of attack that unknown yet. Anomaly-Based may also be referred to as Behavior-based detection [2]. Disadvantages this method is a high false alarm caused by many normal state is not considered as an observable anomaly [7]. Examples of an attacks may be prevented by Anomaly is a DOS-based, intrusion by a legitimate user, Trojan horses. Meanwhile, the SPA (or specification-based) is a technique that identifies the detection deviations from the state protocol by comparing the observed events profiles that has been set. Incompatibility between current behavior with the specifications will be reported as an attack [8]. Unlike with an anomaly that use host-based or network-specific profiles, the SPA relies on vendor profile developers that defines whether the protocol may or may not to be used [8]. Drawback of this method is to consume a huge resource when research and examination of state protocol.

Such mention above, anomaly detection has disadvantages which it obtain high false positive and low detection accuracy. It happened because anomaly detection was conducted by normal behavior in network. In order to detect attack traffic, the system should recognize normal activity. Because of that, a better system for build normal profile is needed. For example using machine learning method. This methods can use to improve the quality of detection attack. The method can be classified into supervised and unsupervised. Many researchers [7–11] are beginning to apply the advance

of machine learning techniques above to address the weaknesses on IDS. Supervised classification is based on the method that concludes a function of labeled training data. This method has a relatively stable system performance and effective to detect known attacks. However, this method is difficult to detect unknown attacks [11, 12], so that the detection accuracy becomes not good. Unknown attack is condition that the type of package of traffic is not identified in the security system. Labeled Data has a high cost and difficult to collect. In addition, this method requires a long time in detecting attacks [13]. Unsupervised learning methods based on clustering and effective to detect this type of unknown attacks [14].

On the other hand, another method was proposed based on hybrid machine learning technique, that is semi-supervised. From studies [11], they are advised to use semi-supervised method to overcome the problem above. Semi-supervised is a amalgamation between supervised and unsupervised. Semi-supervised learning using a large amount of labeled data, and unlabeled data, to build a better classification [7]. In order to improve the IDS problem, Semi-supervised method for anomaly IDS *(SS-IDS)* is proposed. Chen, et al. [10] conducted a study that aims to prove that the semi-supervised method is better than using supervised (Naïve Bayes, Bayesian Networks, Support Vector Machine (SVM), Random Forest, k Nearest Neighbor (kNN), C4.5, and RBF network). They used two classification methods (Spectral Graph Transducer and Gaussian Fields Approach) and one clustering method (MPCK-mean). In [10], the results obtained show that the detection accuracy of the two semi-supervised proposed method is better than a supervised method.

Many researchers use semi-supervised methods to improve detection accuracy. They also use it for comparison against the supervised and unsupervised methods. However, even this method has a shortage, so, semi-supervised become the focus of research due to unsatisfactory performance. In the study [10, 15] The proposed detection accuracy method is still low. On the other hand, Duong and Hai [1] conducted a semi-supervised research model for anomaly detection. But the results of this study indicate that IDS false positives are still high.

## 1.2 Conceptual Framework

The main concept from proposed SS-IDS is to detect an anomaly. Anomaly is unusual behavior. That technique is using normal behavior for basic informations. The research is to increase detection accuracy by improved the quality of training data. By using normal parameter that build in training process, it will help to handle detected a new

traffic with unknown category, whether the data is normal or an attack. The new traffic should be computed to get a parameter and compare with normal parameter.

## 1.3 Problem Statements

In signature detection method based, it only monitors known types of attacks and difficulty to detect unknown attack. Therefore, SS-IDS anomaly based is proposed in order to resolve the problems above. However, Existing SS-IDS methods still has some disadvantages, such as low detection accuracy and high false positives.

## 1.4 Hypothesis

To obtain a high detection accuracy and low false positives, research conducted using SS-IDS to improve the performance of the training phase. The method is the recognition based on incomplete information. When we using big data, it usually has redundant data. To improved the training data and remove the redundant data, filtering multi-collinearity and feature selection is proposed.

## 1.5 Assumption

*Assumption 1*: The majority of the network connections are normal traffic. Only X% of traffic are malicious. (Portnoy, Eskin & Stolfo 2001)

*Assumption 2*: The attack traffic is statistically different from normal traffic. (Javitz & Vadles 1993, Denning 1987)

## 1.6 Scope and Delimitation

The dataset that will be used is NSL-KDD. There are several dataset from NSL-KDD. This research use Train20% both for training and testing. In this study, the focus is to help an administrator to monitor an intrusion.

## 1.7 Importance of The Study

From previous research, IDS has several shortcomings such as low detection accuracy and high false positives. Therefore, the proposed SS-IDS using Outlier Removal Clustering for improve training phase. The uniqueness of the proposed system is used a multicollinearity testing techniques for increasing the performance of the proposed IDS method. This method is used to make a no redundant data. From previous studies [1], they using six attributes that claim can increase accuracy. Therefore, we will proposed

feature selection method that can reduce attributes and improve the performance of the previous method, so as to reduce false positives and improve detection accuracy.