A Novel Metric for the Evaluation of IDSs Effectiveness

Khalid Nasr, Anas El Kalam

A Novel Metric for the Evaluation of IDSs Effectiveness

Khalid Nasr[1] and Anas Abou El Kalam[2]

[1] IRIT, ENSEEIHT-INPT, Université de Toulouse, France
khalid.nasr79@hotmail.com
[2] OSCARS Laboratory, UCA/ENSA of Marrakesh, Morocco
anas@ensa.ac.ma

Abstract. Nowadays intrusion detection system (IDS) has a considerable attention as a crucial element in network security. The question that arises is which IDS is effective for our system? The answer should inevitably take into account the evaluation of IDSs effectiveness. Dealing with this challenge, many valuable evaluation metrics have been introduced such as *receiver operating characteristic* (*ROC*) *curve, Bayesian detection rate, intrusion detection capability, intrusion detection operating characteristic, cost-based metrics*, etc. The benefits and drawbacks of these metrics are discussed in this paper. We subsequently propose a novel metric called *intrusion detection effectiveness* (E_{ID}) that manipulates the drawbacks of the existing ones, taking into account all essential and related parameters. We demonstrate the utility of E_{ID} over the previously proposed ones, and how it realizes the measurement of the actual effectiveness rather than the relative effectiveness as followed by the existing ones. E_{ID} can be used for evaluating the wired or wireless IDSs effectiveness. Additionally, we conduct experimental evaluation of two popular wireless IDSs (WIDSs), *Kismet* and *AirSnare*, to illustrate the benefits of E_{ID}.

Keywords: IDSs effectiveness, evaluation metrics, intrusion detection, false alarms.

1 Introduction

Despite the importance of intrusion detection systems (IDSs) in network security, their performance is sometimes not satisfying in practice. Thus, evaluating the IDSs performance is a pressing necessity. Many attributes judge the IDSs performance such as *effectiveness, efficiency, interoperability* [1], *redundant alerts correlation, attack type recognizing, the impact on the supervised system resources, scalability and flexibility, etc*. No doubt that the IDSs effectiveness is considered the main attribute and basic factor in evaluating the IDSs performance, where it reflects the ability of the IDS to detect the intrusive activities and the absence degree of the false alarms; they are considered the main great challenges facing the IDSs performance.

Evaluation metrics play the significant role in measuring and evaluating the IDSs effectiveness. In this paper, we study the well-known existing metrics for the IDSs effectiveness evaluation, such as *receiver operating characteristic* (*ROC*) [2] [3], *Bayesian detection rate* ($P(I|A)$) [1], *cumulative cost* [4], *expected cost* [5], *intrusion detection capability* (C_{ID}) [6], and *intrusion detection operating characteristic*

(*IDOC*) [7]. Each of these metrics is based on a different theoretical approach such as *decision theory* [5], *information theory* [6], *cost-based analysis* [4] [5], etc. The strengths and weaknesses of these metrics are discussed in this paper, and consequently we propose a novel evaluation metric called *intrusion detection effectiveness* (E_{ID}) that manipulates the drawbacks of the existing ones, especially the common main drawback that is manifested in their main notion of measuring the IDSs effectiveness on the basis of comparing two IDSs or more to select the best one, whereas this selected one may be ineffective.

Our developed metric E_{ID} helps in measuring the actual effectiveness of IDSs rather than measuring the relative effectiveness as followed by the previously proposed metrics. The notion of E_{ID} is based on comparing the operation curve of the IDS under test to the optimal operation curve (i.e., created as a zero reference curve for the optimal operation state) by calculating the variation between the two curves. The variation value interprets the deviation of the IDS operation from the intended optimal operation.

The rest of this paper is organized as follows. Section 2 studies the existing metrics, their benefits, and their drawbacks. Section 3 introduces the novel proposed metric E_{ID} for evaluating the IDSs effectiveness. Section 4 presents the proof of the concept to achieve a credible evaluation of two popular WIDSs (*Kismet* and *AirSnare*). Finally, section 5 presents our conclusion and perspective.

2 Related Work

Various appreciable efforts have been exerted in the recent past for developing reliable evaluation metrics. In this section, we are concerned with analyzing the most valuable and well-known metrics for evaluating the IDSs effectiveness. *Bayesian detection rate* has a great concern in this paper, where we manipulate it to extract the base equation for our proposed metric E_{ID} (*intrusion detection effectiveness*).

Receiver Operating Characteristic (*ROC*). The first unified metric used in the experimental evaluation of IDSs is the *receiver operating characteristic* (*ROC*) curve, as applied by DARPA evaluations [2] [3]. *ROC* curve is used to analyze the trade-off between *detection rate* and *false alarms rate*. The notion of using *ROC* curve in IDSs evaluation is based on comparing the IDSs curves to select the best one. If *ROC* curves don't cross, then the upper curve with the higher values of *detection rate* is considered better than the lower ones. But, if *ROC* curves are crossed, then the differentiation between them is based on the area under each curve. One of the drawbacks of *ROC* curve is its disregard of *base-rate* parameter [1] that is considered a significant parameter in the IDSs evaluation, where it reflects the hostility of the operating environment (i.e., represented by the *prior probability of intrusion*).

Bayesian Detection Rate (*P(I|A)*). *Bayesian detection rate* [1] defines a mathematical relation between the main parameters related to the intrusion detection effectiveness, i.e., *detection rate, false alarms rate,* and *base-rate*. The main advantage of this metric is its consideration of the *base-rate* parameter or probability

of intrusion ($P(I)$). *Bayesian detection rate* ($P(I|A)$) (Eq. 1) was mainly derived from the Bayes' theorem by considering the possible events related to IDSs.

$$P(I|A) = \frac{P(I) \cdot P(A|I)}{P(I) \cdot P(A|I) + P(\neg I) \cdot P(A|\neg I)} \quad \text{Eq. 1}$$

Where I, $\neg I$, A, $P(A|I)$, and $P(A|\neg I)$ denote intrusions, normal traffic, alarms, detection rate, and false alarms rate respectively. Axelsson in [1] studied the effect of the base-rate fallacy on the intrusion detection, and he demonstrated that the limiting factor for IDS performance is not the ability to correctly identify the intrusions, but rather its ability to suppress the false alarms. We totally agree with this conclusion.

Unfortunately, despite the prominence of *Bayesian detection rate* and its consideration of the significant parameters for the IDSs effectiveness evaluation, it is not completely expressive metric for measuring the IDSs effectiveness. We analyze $P(I|A)$ (Eq. 1) mathematically to reach the following results:

Case 1: if $P(A|I) = P(A|\neg I) = 1$

Combining these values with equation Eq. 1, then;

$$P(I|A) = \frac{P(I)}{P(I) + P(\neg I)}$$

Since $P(I) + P(\neg I) = 1$, then;

$$P(I|A) = P(I) \quad \text{Eq. 2}$$

Case 2: if $P(A|I) = 0$

Combining this value and equation Eq. 1, then;

$$P(I|A) = 0 \quad \text{Eq. 3}$$

$P(I|A)$ gives reasonable expressions for the IDSs effectiveness, just in the above two cases. In *case* 1 of passing all the traffic with raised alarms ($P(A|I) = P(A|\neg I) = 1$), $P(I|A)$ equals $P(I)$ (Eq. 2) that is considered the perfect expression in this case; where the ratio of the detected intrusions to the generated alarms corresponds to the ratio of the intrusions to the input traffic. In *case* 2, $P(I|A)$ equals zero when the *detection rate* $P(A|I)$ comes to nought (Eq. 3), where $P(A|I)$ is the predominant parameter in equation Eq. 1. However, the drawback of *Bayesian detection rate* $P(I|A)$ is manifested when the *false alarms* or *false positive rate* $P(A|\neg I)$ is equal or close to zero as shown in the following *case* 3.

Case 3*:* as $P(A|\neg I)$ approaches 0, then equation Eq. 1 can be written as;

$$\lim_{P(A|\neg I)\to 0} P(I|A) = \lim_{P(A|\neg I)\to 0} \frac{P(I) \cdot P(A|I)}{P(I) \cdot P(A|I) + P(\neg I) \cdot P(A|\neg I)} = 1 \quad \text{Eq. 4}$$

As we observe from equation Eq. 4, when $P(A|\neg I)$ equals or approaches "0", $P(I|A)$ is equal to constant value "1" for any value of $P(A|I)$; this unfortunately seems illogical. How the IDSs effectiveness can be evaluated in disregard of the *detection rate* $P(A|I)$? Merely considering the complete absence of false alarms is insufficient. Thus $P(I|A)$, in this case, is inexpressive metric for the IDSs effectiveness or even the detection rate. However, we propose a reasonable solution

for this drawback through our manipulation of $P(I|A)$ to become completely expressive in a new form called *enhanced Bayesian detection rate* (EBD) (section 3.1). We use EBD as a base equation for our proposed metric E_{ID} (section 3.2).

Cost-Based Metrics. Cost-based metrics analyze the intrusion detection from the perspective of costs. Stolfo et al. [4] proposed *cumulative cost* metric for evaluating the fraud and intrusion detection in the financial information systems. They defined three types of costs; *operational, damage,* and *challenge costs*. *Operational cost* refers to the resources needed to run the IDS. *Damage cost* is the amount of damage caused by the leaked undetected intrusions. *Challenge cost* is the cost of acting upon an intrusion when it is detected. $CumulativeCost(S)$ metric (Eq. 5) is derived by considering the challenge and damage costs on the outcome events of IDSs plus the operational cost to evaluate the IDSs over test set "*S"* of labeled connection "c*"*. One of the drawbacks of this metric is its disregard of the *base-rate* parameter.

$$CumulativeCost(S) = \sum_{c \in S} (Cost(c) + OpCost(c)) \qquad \text{Eq. 5}$$

The second noted metric is *expected cost* metric that was proposed by Gaffney and Ulvila [5] who argued that both *ROC* analysis and *cumulative cost* metric are incomplete metrics. They used decision analysis techniques to combine and extend *ROC* analysis and cost-based analysis to provide *expected cost* metric that considers the *base-rate* parameter. More details about this metric are available in [5].

Intrusion Detection Capability (C_{ID}). Gu et al. [6] proposed an information-theoretic measure of the intrusion detection capability. They depended on the notion of having less uncertainty about the IDS input, given the IDS output. They introduced C_{ID} metric (Eq. 6) as the mutual information between the IDS input X and output Y normalized by the entropy (or self-information) of the input $H(X)$. C_{ID} is maximized by calculating the operating point that minimizes the uncertainty of the input.

$$C_{ID} = \frac{I(X;Y)}{H(X)} = \frac{H(X) - H(X|Y)}{H(X)} \qquad \text{Eq. 6}$$

Where $H(X|Y)$ is the entropy of X given Y. We believe that the notion of C_{ID} for minimizing the uncertainty of the input is inapplicable in the IDSs evaluation domain.

Intrusion Detection Operating Characteristic (*IDOC*). Cardenas et al. [7] used $P(I|A)$ (Eq. 1) and introduced the *intrusion detection operating characteristic* (*IDOC*) as a trade-off curve between the $P(I|A)$ and the *probability of intrusion detection* $P(A|I)$. As a consequence of the dependence of *IDOC* on the *Bayesian detection rate* equation, it carries all its drawbacks.

The common drawback of most existing metrics lies in their main notion of comparing two or more IDSs to select the best one, although this selected one may be ineffective. This is considered a deficient approach that leads to measuring the *relative effectiveness* rather than the *actual effectiveness*. We are concerned with manipulating this drawback and the above mentioned ones.

3 Intrusion Detection Effectiveness (E_{ID})

The logical approach for measuring the *actual effectiveness* is comparing the IDS under test to the optimal operation level (as a reference). We thus propose a new evaluation metric E_{ID} (*intrusion detection effectiveness*) that is based on the notion of comparing the operation curve of the IDS under test to the optimal operation curve (created as a zero reference curve) by calculating the variation between the two curves. The variation value interprets the deviation of the IDS from the intended optimal operation. We believe that the main parameters which the IDS effectiveness depends on are *detection rate, false alarms rate,* and *base-rate*. To realize the notion of E_{ID}, we need an expressive formula or equation that considers these parameters to be used as a base for E_{ID}. As a result of our research, we discovered that *Bayesian* detection rate (Eq. 1) regards the needed parameters, but it is inappropriate as a base equation due to its drawback when the *false alarms* equals or approaches "0" (Eq. 4). Consequently, we manipulate this drawback to derive a new completely expressive formula called *enhanced Bayesian detection rate* (EBD) to become the base for E_{ID}.

3.1 Deriving the Enhanced Bayesian Detection Rate (EBD)

As a brief summary of our analysis of *Bayesian detection rate* $P(I|A)$ (section 2), in *case* 1 ($P(A|I) = P(A|\neg I) = 1$) and *case* 2 ($P(A|I) = 0$), $P(I|A)$ gives reasonable expressions, but it is inexpressive in *case* 3 (as $P(A|\neg I)$ equals or approaches "0"). Accordingly, we are concerned with manipulating *case* 3. By analyzing *case* 3, we conclude that the logical expressive formula for $P(I|A)$, as $P(A|\neg I)$ equals or approaches "0", should be equal to $P(A|I)$. This can be achieved by modifying the denominator of equation Eq. 4 to produce the following new formula.

$$\lim_{P(A|\neg I)\to 0} P(I|A) = \lim_{P(A|\neg I)\to 0} \frac{P(I)\cdot P(A|I)}{P(I)\cdot P(A|I)_{\rightsquigarrow=1} + P(\neg I)\cdot P(A|\neg I)}$$
$$= \frac{\cancel{P(I)}\cdot P(A|I)}{\cancel{P(I)}\cdot P(A|I)_{\rightsquigarrow=1}} = P(A|I) \qquad \text{Eq. 7}$$

From equations Eq. 1, Eq. 2, Eq. 3 and Eq. 7 we produce the *enhanced Bayesian detection rate* (EBD) (Eq. 8) that is completely expressive under all operation conditions.

$$EBD = \frac{P(I)\cdot P(A|I)}{P(I) + P(\neg I)\cdot P(A|\neg I)} \qquad \text{Eq. 8}$$

Property 1

EBD can be defined as the posterior probability of detected intrusion (TP) given the total output of intrusion related responses ($TP + FN$) and false alarms (FP).

Proof

The intrusion detection can be summarized by the simple model shown in Fig. 1, where I, $\neg I$, A, $\neg A$, TP, FP, FN, and TN denote intrusions, normal traffic, alarms, no alarms, true positives, false positives, false negatives, and true negatives respectively.

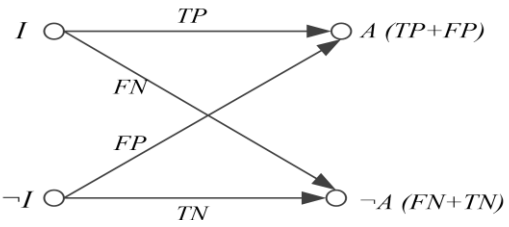

Fig. 1. Intrusion Detection Model.

Basically,

$$P(I) = I/(\neg I + I)$$
$$P(\neg I) = \neg I/(\neg I + I)$$
$$P(A|I) = TP/I = TP/(TP + FN)$$
$$P(A|\neg I) = FP/\neg I = FP/(FP + TN)$$

Recalling equation Eq. 8 and solving it by these parameters, then;

$$EBD = \frac{P(I)\cdot P(A|I)}{P(I) + P(\neg I)\cdot P(A|\neg I)} = \frac{1}{\frac{1}{P(A|I)} + \frac{P(\neg I)\cdot P(A|\neg I)}{P(I)\cdot P(A|I)}}$$
$$= \frac{1}{\frac{TP+FN}{TP} + \frac{\cancel{\neg I}\cdot FP/\cancel{(FP+TN)}}{\cancel{I}\cdot TP/\cancel{(TP+FN)}}} = \frac{TP}{TP+FN+FP} \qquad \text{Eq. 9}$$

Equation Eq. 9 shows the significance of EBD for measuring the proportion of the intrusion related responses ($TP + FN$) and false alarms (FP) that correspond to the detected intrusions (TP). This is considered one of the great advantages of EBD over $P(I|A)$ that ignores the *false negatives* (*FN*), as demonstrated in the following.

By recalling $P(I|A)$ (Eq. 1) and solving it in the same way, then;

$$P(I|A) = \frac{P(I)\cdot P(A|I)}{P(I)\cdot P(A|I) + P(\neg I)\cdot P(A|\neg I)} = \frac{1}{1 + \frac{P(\neg I)\cdot P(A|\neg I)}{P(I)\cdot P(A|I)}}$$
$$= \frac{1}{1 + \frac{\cancel{\neg I}\cdot FP/\cancel{(FP+TN)}}{\cancel{I}\cdot TP/\cancel{(TP+FN)}}} = \frac{TP}{TP+FP} \qquad \text{Eq. 10}$$

Equation Eq. 10 clarifies that $P(I|A)$ disregards *FN* parameter that influences the expressiveness of the metric.

To clarify more the benefit of EBD over $P(I|A)$ regarding taking the whole false responses (*FP*+*FN*) into account, the relationships between the IDS input and output events are depicted through Venn diagram (Fig. 2). The intersections between the different events are represented by the areas denoted by numbers from 1 to 4. Area 1 represents the tranquil area of no intrusion and no alarm, but areas 2, 3 and 4 represent the challenge areas of false responses (areas 2 and 4) and detected intrusions (area 3). These events in the areas 2, 3 and 4 have a great significance in the IDSs effectiveness evaluation, and they should be considered by the evaluation metric. This is attained by EBD as shown in equation Eq. 9. On the contrary, $P(I|A)$ considers only the events of areas 3 and 4 as shown by Eq. 10.

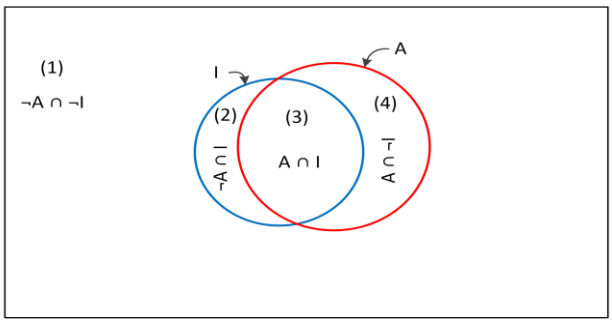

Fig. 2. The Relationships between the IDS Input and Output Events.

Property 2

EBD is an expressive metric under different operation conditions.

Proof

EBD agrees with $P(I|A)$ on the aforementioned *case* 1 (Eq. 2) and *case* 2 (Eq.3), and gives the same expressive results. As well, *EBD* gives an expressive value in the third case (Eq. 7) according to its mathematical manipulation. For the other rest conditions, *EBD* can be analyzed as follows.

Case 4**:** if $P(A|\neg I) = 1$

Combining this value with equation Eq. 8, then;

$$EBD = \frac{P(I) \cdot P(A|I)}{P(I) + P(\neg I)}$$

Since $P(I) + P(\neg I) = 1$, then;

$$EBD = P(I) \cdot P(A|I) = \frac{\cancel{I}}{(\neg I + I)} \cdot \frac{TP}{\cancel{I}} = \frac{TP}{TP + FN + FP} \qquad \text{Eq. 11}$$

Equation Eq. 11 demonstrates the expressiveness of EBD in the worst case of false alarms ($P(A|\neg I) = 1$).

Case 5**:** if $P(A|I) = 1$

Combining this value with equation Eq. 8, then;

$$EBD = \frac{P(I)}{P(I) + P(\neg I) \cdot P(A|\neg I)} = \frac{1}{1 + \frac{\cancel{\neg I} \cdot FP / \cancel{(FP + TN)}}{TP}} = \frac{TP}{TP + FP} \qquad \text{Eq. 12}$$

As shown in equation Eq. 12, EBD equals the proportion of the generated alarms that correspond to the detected intrusions. This is the expressive formula in this case of the absence of false negatives ($P(A|I) = 1 \Leftrightarrow FN = 0$). The above analysis concludes the expressiveness of EBD under different operation conditions.

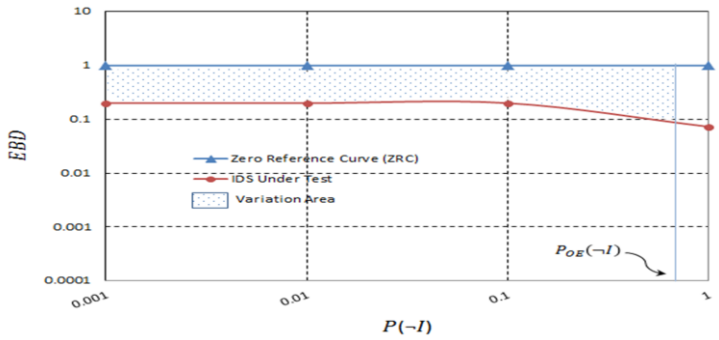

Fig. 3. The trade-off between *EBD* and *P(¬I)*.

3.2 Deriving the Intrusion Detection Effectiveness (E_{ID})

Following the main notion of our metric E_{ID}, we consider the trade-off between *EBD* and $P(\neg I)$ that helps in developing the expressive metric E_{ID}. To simplify dealing with *EBD* (Eq. 8), we adapt it to be a function of $P(\neg I)$ as shown in Eq. 13.

$$EBD = \frac{P(A|I) - P(\neg I) \cdot P(A|I)}{1 - P(\neg I) \cdot (1 - P(A|\neg I))} \qquad \text{Eq. 13}$$

The first step in deriving E_{ID} is calculating and plotting the *zero reference curve* (*ZRC*) as a trade-off between *EBD* and $P(\neg I)$ with assumption of the optimal operation case of the IDS under test. To clarify the idea of calculating and plotting *ZRC,* we assume an IDS installed in an operating environment with hostility or probability of intrusion $P(I)$=3*10^{-4}. Then, the probability of no intrusion $P(\neg I) = 1 - P(I) = 0.9997$. First, we assume that the IDS under test operates at the optimal case with perfect *detection rate* ($P(A|I)$=1) and complete absence of *false alarms* ($P(A|\neg I) = 0$). We combine these values with *EBD* (Eq. 13) to plot *ZRC* (Fig. 3). As a note, the axes are set to logarithmic scale. The second step, we plot the real operation curve of the IDS with the actual values of the *detection rate* and *false alarms*; we assume their values as $P(A|I) = 0.2$ and $P(A|\neg I) = 0.0035$. Now we have two operation curves; one as a *ZRC* curve for the optimal operation and the other represents the actual operation curve (Fig. 3). The variation between the two curves is represented by the dotted area. $P_{OE}(\neg I)$ denotes the probability of no intrusion in the operating environment, and it refers to the upper limit of the variation area.

We normalize this variation by the area under *ZRC* (only through $P(\neg I) = [0, P_{OE}(\neg I)]$) to have a representative metric E_{ID} of values in the range [0, 1]; where "0" indicates zero deviation from the intended optimal operation and then perfect effectiveness, but "1" indicates the maximum deviation and then zero effectiveness. E_{ID} is represented by equation Eq. 14, where EBD_{ZRC}, $P_{ZRC}(A|I)$, and $P_{ZRC}(A|\neg I)$ denote *EBD*, detection rate, and false alarms of *ZRC* respectively. As well, EBD_{ID}, $P_{ID}(A|I)$, and $P_{ID}(A|\neg I)$ denote *EBD*, detection rate, and false alarms of IDS under test respectively. $P(\neg I)$ is considered the integration variable.

$$E_{ID} = \frac{1}{\int_0^{P_{OE}(\neg I)} EBD_{ZRC}\ dP(\neg I)} \left(\int_0^{P_{OE}(\neg I)} EBD_{ZRC}\ dP(\neg I) - \int_0^{P_{OE}(\neg I)} EBD_{ID}\ dP(\neg I) \right) \quad \text{Eq. 14}$$

Where,

$$EBD_{ZRC} = \frac{P(I) \cdot P_{ZRC}(A|I)_{\rightsquigarrow=1}}{P(I) + P(\neg I) \cdot P_{ZRC}(A|\neg I)_{\rightsquigarrow=0}} = 1 \quad \text{Eq. 15}$$

$$EBD_{ID} = \frac{P_{ID}(A|I) - P(\neg I) \cdot P_{ID}(A|I)}{1 - P(\neg I) \cdot (1 - P_{ID}(A|\neg I))} \quad \text{Eq. 16}$$

Then equation Eq. 14 becomes;

$$E_{ID} = 1 - \frac{\int_0^{P_{OE}(\neg I)} EBD_{ID}\ dP(\neg I)}{P_{OE}(\neg I)} \quad \text{Eq. 17}$$

Equation Eq. 17 can be solved mathematically [8], and E_{ID} becomes;

$$E_{ID} = 1 - \frac{P_{ID}(A|I) \cdot \left(P_{OE}(\neg I) + \left(\frac{1}{(1 - P_{ID}(A|\neg I))} - 1 \right) ln\left|1 - P_{OE}(\neg I) \cdot (1 - P_{ID}(A|\neg I))\right| \right)}{P_{OE}(\neg I) \cdot (1 - P_{ID}(A|\neg I))} \quad \text{Eq. 18}$$

Property 3

E_{ID} is an expressive metric for measuring the actual effectiveness of IDSs by values in the range [0,1], where "0" indicates the ideal case of supreme effectiveness, but "1" indicates the worst case of zero effectiveness.

Proof

Case 1: when the IDS detects all the intrusive activities ($P(A|I) = 1$) and generates no false alarms ($P(A|\neg I) = 0$), then its deviation from the optimal operation case can be measured by E_{ID} (Eq. 18) as follows.

$$E_{ID} = 1 - \frac{P_{ID}(A|I)_{\rightsquigarrow=1} \cdot (P_{OE}(\neg I) + 0)}{P_{OE}(\neg I) \cdot (1 - P_{ID}(A|\neg I)_{\rightsquigarrow=0})} = 1 - \frac{\cancel{P_{OE}(\neg I)}}{\cancel{P_{OE}(\neg I)}} = 0 \quad \text{Eq. 19}$$

Equation Eq. 19 demonstrates the supreme effectiveness of the IDS by its zero deviation from the optimal operation case.

Case 2: when the IDS fails to detect the intrusions ($P(A|I) = 0$), its deviation from the optimal operation case can be measured by E_{ID} (Eq. 18) as follows.

$$E_{ID} = 1 - \frac{0}{P_{OE}(\neg I) \cdot (1 - P_{ID}(A|\neg I))} = 1 \quad \text{Eq. 20}$$

Equation Eq. 20 demonstrates the maximum deviation of the IDS from the optimal operation case and accordingly its ineffectiveness.

Besides *property* 1 and *property* 2 of *EBD* that is the base equation of E_{ID}, it becomes clear from equations Eq. 19 and Eq. 20 that E_{ID} is an expressive metric for measuring the actual effectiveness. E_{ID} can be used for evaluating the effectiveness of wired or wireless IDSs.

4 Proof of the Concept

As a proof of the concept, we conduct an experimental evaluation of two popular wireless IDSs (WIDSs) *Kismet* (for Linux) [9] and *AirSnare* (for Windows) [10], and measure their effectiveness using E_{ID}. We used RF shielded testbed, an access point Linksys WRT54GL, and workstations (Linux and Windows) with Wi-Fi adapters ALFA awus036h and D-Link DWA-110 with respect to the compatibility with the operating systems and the WIDSs. We resorted to use the RF shielded testbed to circumvent the problem of the uncontrolled 802.11 traffic from the adjacent wireless stations that obstructs the accurate measurements of the considered parameters. We are concerned with the wireless infrastructure mode with two possible scenarios for the installation of WIDSs (Fig. 4); *scenario* 1: the WIDS was installed on the access point, and *scenario* 2: the WIDS was installed on a terminal machine as a victim. As for the normal background traffic, we generated real traffic by capturing the operational traffic during the normal operation of a private network installed for this purpose, and then replaying the collected traffic into the testbed. This private network (Fig. 4) consists of an access point, three workstations (i.e., two machines operate under Windows and the third one operates under Linux) and two mobile phones (i.e., Android system). Table 1 shows the statistics of the collected benign traffic.

As for the generated attacks, the credible evaluation of WIDSs necessitates taking into account all possible attacks. While this is operationally impossible, it is necessary to select representative attack test cases that are extracted mainly from a holistic classification of wireless attacks. Dealing with this challenge, we used our developed taxonomy of wireless attacks from the perspective of the WIDSs evaluator [11] and we generated the attacks listed in Table 2 according to the representative attack test cases shown in Fig. 5. As well, the attack detection of each WIDS is shown in Table 2. For calculating the detection rate, if we follow the ordinary method that was used in the previous evaluations of IDSs, then Table 2 is sufficient for the calculations and then the detection rate is $P(A|I) = 0.61$ for *Kismet*, and $P(A|I) = 0.167$ for *AirSnare*. These values are not real expressive values for the detection rate, and subsequently have a negative effect on the calculation of the real effectiveness. The best way for calculating the expressive detection rate is considering the probability of occurrence of the attack test cases under the operating environment conditions.

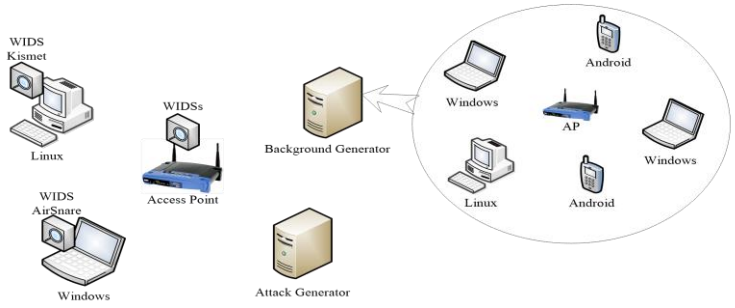

Fig. 4. Evaluation Testbed.

Table 1. The Collected Normal Traffic.

Frame Subtype	Frame Count
Association request	38
Association response	43
Reassociation request	172
Reassociation response	142
Probe request	227081
Probe response	218602
Beacon	378
Disassociation	332
Authentication	169
Deauthentication	80
Action frames	3484
Null data	89920
QoS data	1723
QoS null data	19868
Total	562032

Table 2. The Attack Detection.

Generated Attacks	Kismet	AirSnare
	TP	*TP*
Deauthentication/Disassociation Flood (< 10 Request)	✓	x
Deauthentication/Disassociation Flood (< 20 Request)	✓	x
Deauthentication/Disassociation Flood (> 30 Request)	✓	✓
Deauthentication/Disassociation Flood (> 100 Request)	✓	✓
Deauthentication/Disassociation (Amok mode)	✓	x
Fake Authentication	✓	x
Authentication Flood	✓	x
Beacon Flood (evil duplicate AP DoS)	✓	x
MITM attack	x	x
ARP Request Replay Attack	x	x
WPA Downgrade	✓	x
WPA Cracking	✓	x
WEP Cracking	x	x
Chopchop	x	x
Hidden SSID Brute Force	x	x
Rogue AP	x	x
RF Jamming	✓	x
MAC Spoofing	x	✓

In our evaluation tests, we considered and used 100 attack instances of the attacks listed in Table 2. We considered the instances of the generated attacks by ratios that correspond approximately to the probability of attack occurrence in some real systems. This consideration was managed according to our statistical analysis of the registered wireless attacks and vulnerabilities in the popular database such as Common Vulnerabilities and Exposures [12], National Vulnerability Database –

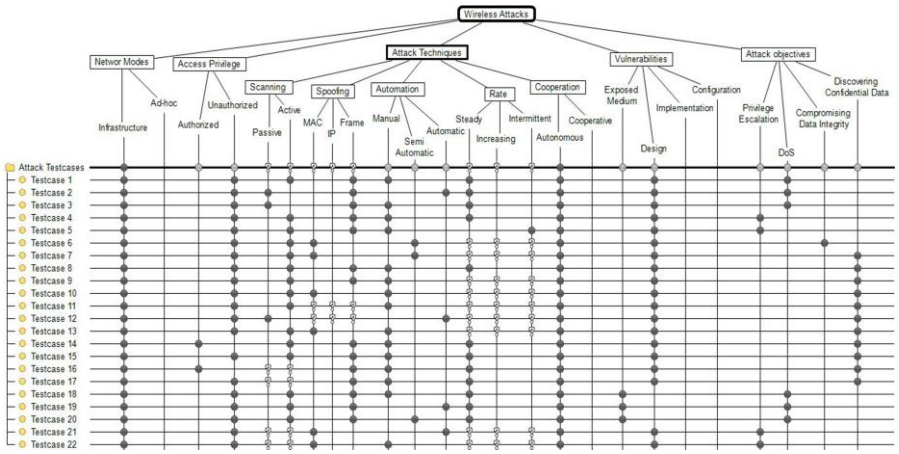

Fig. 5. Representative Attack Test Cases.

Table 3. The Generated Attacks and the Corresponding Test Cases.

Generated Attacks	Representative Attack Test Cases
Deauthentication/Disassociation Flood	1, 2, 3
Deauthentication/Disassociation (Amok mode)	3
Fake Authentication	4, 5
Authentication Flood	2,3
Beacon Flood (evil duplicate AP DoS)	3
MITM attack	6,7
ARP Request Replay Attack	8
WPA Downgrade	3
WPA Cracking	9
WEP Cracking	10, 11, 12
Chopchop	13
Hidden SSID Brute Force	11
Rogue AP	14, 15, 16, 17
RF Jamming	18, 19, 20
MAC Spoofing	21, 22

NIST [13], and others. It is worth mentioning that we considered in our calculations the deauthentication/disassociation flood attack instances with deauthentication requests > 30 (Table 2); we generated it by 8 instances from total of 100 instances of all the generated attacks. We classified the generated attacks under the representative test cases (Fig. 5 and Table 3), and adjusted the estimated probability of occurrence as shown in Table 4. Then, the expressive detection rate is $P(A|I) = 0.65$ for *Kismet*, and $P(A|I) = 0.13$ for *AirSnare*. In our evaluation environment, the used 100 attack instances generated approximately 1500 malicious frames, in addition to the generated background normal traffic (Table 1). Then, we have hostility or intrusion probability $P(I)$=1500/562032=2.66889*10^{-3}, and no intrusion probability $P(\neg I)$=0.99733. As well, the registered false alarms for the two WIDSs are

Table 4. Probability of Occurrence of the Generated Attack Instances.

Attack Test Cases	Estimated Probability	WIDSs Detection Ratio	
		Kismet	AirSnare
1, 2, 3, 4, 5	0.46	✓	✓ (0.08)
6, 7	0.05	x	x
8	0.04	x	x
9	0.13	✓	x
10, 11, 12	0.1	x	x
13	0.04	x	x
14, 15, 16, 17	0.07	x	x
18, 19, 20	0.06	✓	x
21, 22	0.05	x	✓
Total	1	0.65	0.13

$P(A|\neg I)$=0.008967 for *Kismet* and $P(A|\neg I)$=0.0014946 for *AirSnare*. Combining these obtained results with our proposed evaluation metric E_{ID} (Eq. 18), then;

$$E_{ID}\ (kismet) = 0.37$$

$$E_{ID}\ (AirSnare) = 0.871$$

In the same way, as described in section 3.2, we can plot the operation curves of the two WIDSs, besides the *zero reference curve* (*ZRC*) as shown in Fig. 6. E_{ID} of the two WIDSs and Fig. 6 show that *Kismet* operation doesn't deviate much more from the optimal case *ZRC*, in contrast to *AirSnare* that has a great deviation from the optimal case. Then, *Kismet* is more effective than *AirSnare*.

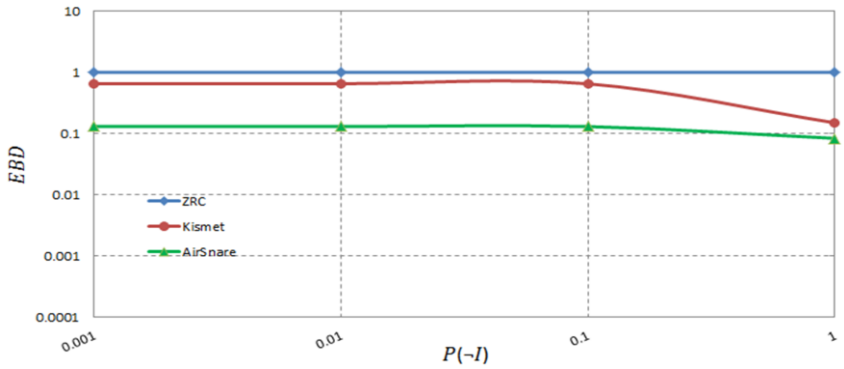

Fig. 6. The Trade-off between *EBD* and *P(¬I)* of *Kismet*, *AirSnare*, and *ZRC*.

5 Conclusion

Our proposed metric E_{ID} manipulated the drawbacks of the existing metrics and it realizes the measurement of the actual effectiveness, taking into account the main related parameters. We conducted credible evaluation of two popular WIDSs (*Kismet*

and *AirSnare*) using E_{ID} and considered some important aspects that were ignored in the existing work such as the probability of occurrence of attacks and selecting the attacks on the basis of representative attack test cases. The results demonstrated that *Kismet* is more effective than *AirSnare*. We are interested in deriving other evaluation metrics for the rest attributes of IDSs/WIDSs performance.

References

1. Axelsson, S.: The Base-Rate Fallacy and its Implications for the Difficulty of Intrusion Detection. Proceedings of the 6th ACM Conference on Computer and Communications Security (CCS'99). pp. 1–7. ACM Press (1999).
2. Lippmann, R., Fried, D., Graf, I., Haines, J., Kendall, K., McClung, D., Weber, D., Webster, S., Wyschogrod, D., Cunningham, R., Zissman, M.: Evaluating Intrusion Detection Systems: The 1998 DARPA Off-line Intrusion Detection Evaluation. Proceedings of the DARPA Information Survivability Conference and Exposition (DISCEX'00). pp. 12–26. , Los Alamitos, CA, USA (2000).
3. Lippmann, R., Haines, J.W., Fried, D., Korba, J., Das, K.: Analysis and Results of the 1999 DARPA Off-Line Intrusion Detection Evaluation. Proceedings of International Symposium on Recent Advances in Intrusion Detection (RAID'00). pp. 162–182. Springer, LNCS 1907, Toulouse, France (2000).
4. Stolfo, S., Fan, W., Lee, W., Prodromidis, A., Chan, P.: Cost-based modeling for fraud and intrusion detection: Results from the JAM project. Proceedings of DARPA Information Survivability Conference and Exposition (DISCEX'00). pp. 130–144 (2000).
5. Gaffney, J.E., Ulvila, J.W.: Evaluation of intrusion detectors: A decision theory approach. Proceedings of the IEEE Symposium on Security and Privacy (S&P'01), Oakland, CA, USA. pp. 50–61 (2001).
6. Gu, G., Fogla, P., Dagon, D., Lee, W., Skoric, B.: Measuring Intrusion Detection Capability: An Information-Theoretic Approach. Proceedings of ACM Symposium on InformAtion, Computer and Communications Security (ASIACCS'06), Taipei, Taiwan (2006).
7. Cardenas, A.A., Baras, J.S., Seamon, K.: A Framework for the Evaluation of Intrusion Detection Systems. Proceedings of IEEE Symposium on Security and Privacy (S&P'06). pp. 63–77 (2006).
8. Zill, D.G., Wright, W.S.: Advanced Engineering Mathematics. Jones & Bartlett Learning (2012).
9. Kismet: Kismet_WIDS, http://www.kismetwireless.net/.
10. AirSnare: AirSnare-WIDS, http://home.comcast.net/~jay.deboer/airsnare/.
11. Nasr, K., Abou El Kalam, A., Fraboul, C.: An IDS Evaluation-Centric Taxonomy of Wireless Security Attacks. International Conference on Network Security & Applications (CNSA'11). pp. 402–413, vol. 196. Springer, CCIS Series, Chennai, India (2011).
12. CVE Wireless: Common Vulnerabilities and Exposures - Wireless, http://cve.mitre.org/cgi-bin/cvekey.cgi?keyword=wireless.
13. NVD-NIST: National Vulnerability Database. National Institute of Standards and Technology (NIST), http://nvd.nist.gov/home.cfm.