CHALMERS
UNIVERSITY OF TECHNOLOGY

Towards an information-theoretic framework of intrusion detection for composed systems and robustness analyses

Computers & Security 116 (2022) 102633

Contents lists available at ScienceDirect

Computers & Security

journal homepage: www.elsevier.com/locate/cose

Towards an information-theoretic framework of intrusion detection for composed systems and robustness analyses

Tobias Mages[a,*], Magnus Almgren[b], Christian Rohner[a]

[a] Department of Information Technology, Uppsala University, Uppsala 752 36, Sweden
[b] Department of Computer Science and Engineering, Chalmers University of Technology, Gothenburg 412 96, Sweden

ARTICLE INFO

Article history:
Received 29 June 2021
Revised 1 December 2021
Accepted 29 January 2022
Available online 2 February 2022

Keywords:
Network intrusion detection
Adversarial robustness
Data-driven evaluation approaches
Performance evaluation metrics
Information theoretic framework
Composed detection systems

ABSTRACT

Network-based Intrusion Detection Systems (NIDSs) are an important mechanism to identify malicious behaviour or policy violations within a network. Such detection systems typically face several challenges, among which are the base-rate fallacy and the resilience against adaptive adversaries. These challenges are often countered in modern NIDSs by combining multiple detection systems to diversify the used feature levels or utilize the advantages of multiple detection methods. However, currently there exists no suitable framework for a detailed analysis of such composed systems. Therefore, the contribution of this work is an evaluation framework for composed systems, which builds on previous information-theoretic approaches and highlights the utility of information-theoretic redundancies for robustness evaluations. This framework enables an attribution of the overall system performance to its individual components, to fine-tune parameters and to study the dynamics between classifiers. The versatility of the framework is demonstrated by designing and evaluating a composed NIDS example based on systems described in the literature and using an open data set. Studying the impact of an evasion attempt with adversarial examples on this system highlighted the importance of robustness against false-alarms as well as detection evasion. Moreover, the framework enables general insights on how to improve the design of composed NIDSs: based on the dynamics between classifiers, it can be shown that optimizing the operation point of each component individually does not necessarily maximize the overall system performance from an information-theoretic perspective. Additionally, it can be shown that existing classification redundancies might not be fully utilized during an attack on the NIDS components, due to a static system design.

1. Introduction

Network-based Intrusion Detection Systems (NIDSs) are an important mechanism to identify malicious behaviour or policy violations within a network. The evaluation of NIDSs is tightly related to their design and can focus on a wide variety of aspects, like the system requirements during operation, its scalability to different traffic loads, the ability to detect new attack types or the adaptability to changes over time (Axelsson, 1999; Mell et al., 2003). This work focuses on the evaluation of the classification performance, which is often data-driven and relates to a number of design challenges.

One key challenge when designing a NIDS is the base-rate fallacy (Axelsson, 1999). The base-rate p refers to the probability of an intrusion $P(I)$ and therefore the ratio of attack samples in the test set, while $P(A)$ refers to the probability of an alarm. For a NIDS, it is desirable to achieve a high True-Positive Rate (TPR, $P(A|I)$) and a high Positive Predictive Value (PPV, $P(I|A)$) in the operation environment. Axelsson (1999) demonstrated the issue of the base-rate fallacy, where the False-Alarm Rate (FPR, $P(A|\neg I)$) limits the overall system performance due to the high class imbalance at base-rates such as 10^{-3} to 10^{-7}.

The second key challenge when designing a NIDS is the resilience against adaptive adversaries. Similarly to how a sample perturbation might be used to evade a matching signature, adversarial examples might be used to evade learning-based detection methods. Adversarial examples aim on finding small perturbations to an input sample, which would lead to a misclassification by the detection system. While this research area has been mainly driven

* Corresponding author.
E-mail addresses: tobias.mages@it.uu.se (T. Mages), magnus.almgren@chalmers.se (M. Almgren), christian.rohner@it.uu.se (C. Rohner).

https://doi.org/10.1016/j.cose.2022.102633

by the image processing domain (Carlini and Wagner, 2017; Goodfellow et al., 2015; Papernot et al., 2016), the techniques are also increasingly applied to constrained domains such as network traffic. Several works already demonstrated that these methods can be adopted for network traffic to evade learning-based NIDSs and highlighted that adaptive adversaries should be considered in system evaluations (Hartl et al., 2020; Hashemi et al., 2019; Sheatsley et al., 2020).

Due to these design challenges, modern NIDSs are often composed of several detection methods or systems. This can enable a reduction of false alarms at low base-rates (Meng and Kwok, 2013), diversify the used feature levels to increase robustness or enable utilizing the advantages of different detection methods (Gu et al., 2008). In this work we consider *composed systems* constructed by chaining classifiers. This makes composed systems a subset of ensemble classifiers which restrict the aggregation function (1) to be expressible in boolean algebra and (2) to operate on the categorical output of the individual detectors. However, the design of such systems is challenging. It is not obvious how the integration of an additional classifier would affect the overall risk of evasion or which component is limiting the system performance. Unfortunately, and to the best of our knowledge, currently there exists no framework that would be suitable for a detailed analysis of these systems in the area of NIDSs.

We consider the *performance* of a detection system to be defined by the uncertainty reduction about the class of a sample that is achieved from knowing the detection outcome, which leads to the mutual-information as performance metric (Gu et al., 2006a). This approach allows us to consider the robustness of composed systems from two perspectives: the classifier robustness and the system robustness. The robustness of a *classifier* is determined by its sensitivity to evasion and thus by its performance loss during an adversarial attack. The robustness of the *system* reflects its ability to compensate for the performance degradation of one of its components, which we explore by evaluating the information-theoretic redundancies within the system.

The main contributions of this work are (1) an Information-Theoretic Framework (ITF) for the analysis of composed NIDSs, which (2) enables the use of information-theoretic redundancies for evaluating the robustness of composed systems. Our framework builds on the work of Gu et al. (2006b), which introduced the information-theoretic approach for the analysis of an individual NIDS.

The proposed framework of this work enables an analysis and comparison of composed systems by quantifying the performance of their individual components, including multiple feature representations, detection methods and their specific arrangement. It can be used to study the dynamics between operation points to fine-tune parameters or evaluate threat models and attack methods by analyzing the robustness dependencies between different classifiers.

The versatility of the framework is demonstrated by applying it to guide the design, fine-tuning and evaluation of a composed NIDS example based on systems described in the literature and study the impact of an evasion attempt with adversarial examples. In particular we attribute the overall system performance to its individual components, show the impact of compositions on their operation points and make statements about the system performance at different base rates. Additionally, the analysis provides general insights on how to improve the design of robust composed NIDSs. The results indicate that an independent operation point optimizations for each component does not maximize the overall system performance from an information-theoretic perspective and that composed systems can contain classification redundancies which might not be fully utilized during an evasion attempt due to the system design.

2. Background

Analyzing a composed NIDS requires the comparable evaluation of individual components or systems. Therefore, accepted evaluation methods and metrics of the area will be discussed to identify the requirements on an analysis framework. This also highlights which issues the previously proposed information-theoretic approaches solved and provides the required background information for its further extension. Finally, methods for evaluating NIDS robustness will be discussed to understand which insights the proposed framework of this work can provide to improve the design of robust systems.

2.1. Evaluation units and metrics

To better understand what an analysis framework and metric should provide in the area, requirements for comparable system evaluations will be discussed together with which limitations existing metrics, both trade-off based and combined optimization objectives, provide in different use cases.

The evaluation of a NIDS can focus on a wide variety of aspects, like for example the coverage of different attack classes, its ease of use, interoperability, transparency or explainability, the adaptability to changes over time, its scalability to different traffic loads, the ability to detect new attack types or the system requirements during operation (Axelsson, 1999; Mell et al., 2003). This work focuses exclusively on a data-driven performance evaluation of the detection system output and therefore its classification performance.

NIDSs typically utilize features of three distinct classes (Giacinto and Roli, 2002; Lee and Stolfo, 2000), which could restrict the system's unit of analysis: Intrinsic features (packet-level) are directly extractable from an individual flow (e.g. duration, flags, used protocol). Traffic features (flow-level) are based on aggregated flow information or statistical information related to past connections, while content features are based on the payload information. Additionally, physical-layer features could be considered as a fourth class to utilize properties of the transceiver, channel or environment (Birnbach et al., 2019; Jiang et al., 2013; Yan et al., 2020).

Comparing systems with different unit of analysis is problematic as the unit of analysis can affect the base-rate and results of an evaluation. Therefore, system comparisons should be based on the same unit of analysis which might require a conversion from the system's unit of analysis to the desired unit of analysis. As example, Gu et al. (2006a) proposed a conversion from packet- to flow-level by defining a flow as malicious if it contains at least one malicious packet. The approach can similarly be applied for converting the results of physical-layer detection systems to flow-level. Assuming that a suitable data set with all required features existed, then these conversions enable a comparable evaluation of different detection systems or enable analyzing composed detection systems (see Section 3.2) which might utilize a variety of detection units.

Besides having a specific unit of analysis, many detection methods can operate on a range of detection thresholds, which are also known as operation points. This leads to the question of which detection threshold should be used during deployment and by evaluating the system performance. Therefore, Cardenas et al. (2006) highlighted that the performance evaluation of NIDSs can be viewed as a multi-criteria optimization problem of maximizing the TPR and FPR or PPV and Negative Predictive Value (NPV). Such issues can be approached either by evaluating trade-off curves, which will be discussed first, or by combining the different criteria into a single optimization objective, which will be discussed afterwards.

One method for analyzing the classification performance of a NIDS is the Receiver Operating Characteristic (ROC), which visualizes the trade-off between the TPR and FPR for different operation points (Hancock, 1966). The ROC curve can be used to compare two detection systems. However, the result is only conclusive if one of the systems performs better for all operation points (the ROC-curves do not intersect). It has also been criticized for being base-rate independent, which does not enable capturing the issue of the base-rate fallacy (Nasr and El Kalam, 2014). Therefore it is important to compare ROC-curves in meaningful false alarm ranges to avoid misleading results.

To present the properties of interest, a Precision Recall Curve (PRC) could be used instead, which is also known as Intrusion Detector Operating Characteristic (IDOC) (Cardenas et al., 2006). It visualizes the trade-off between the PPV and TPR and is typically shown for several base-rates. While it avoids misleading results due to the base-rate fallacy, it maintains the drawbacks of a trade-off curve with being inconclusive in comparisons if two systems intersect and being unable to recommend an ideal operation point.

To overcome the issues of a trade-off based analysis, optimization objectives that combine the different criteria can be defined. A possible approach would be to view the Area Under the ROC (AUC) (Durst et al., 1999), the area of a PRC or the area of a PPV/FPR-curve (Nasr et al., 2012). However, each of these measures would consider multiple operation points during the evaluation, which is not accepted in the community since it would practically be fine-tuned to a specific detection threshold (Gu et al., 2006a; Milenkoski et al., 2015). Therefore, typical comparisons are based on the best operation point or best worst-case performance (Gu et al., 2006a). One metric which avoids the issue of multiple operation points is the Intrusion Detection Effectiveness (E_{ID}), which evaluates the area of a Critical Sucess Index (CSI)/base-rate curve (Nasr and El Kalam, 2014). The definition of E_{ID} considers however base-rates from a specified threshold up to one, such that the impact of high class imbalances in deployment environments may not be reflected in the evaluation result.

Gaffney and Ulvila (2001) highlighted that it should be considered to adopt the evaluation to the specific environment of interest in terms of the assumed cost of a false alarm (c_α) and cost of a missed intrusion (c_β). This leads to a cost ratio $c = c_\beta / c_\alpha$ and two possible optimization objectives. They consider that an operator can act contrary to the detection result, which leads to the optimization objective: $c_{op} = \min(c\beta p, (1-\alpha)(1-p)) + \min(c(1-\beta)p, \alpha(1-p))$, where p is the base-rate, α the FPR and β the False-Negative Rate (FNR). Gu et al. (2006a) demonstrated however that this measure may become independent of the TPR and FPR depending on the selected cost ratio. This issue can be avoided with the *expected cost*, which assumes that the operator does not act contrary to the detection result: $c_{exp} = c\beta p + \alpha(1-p)$ (Gaffney and Ulvila, 2001; Meng, 2012).

Gu et al. (2008) studied the problem of alarm fusion for systems of multiple classifiers from a cost perspective. They highlight that the Likelihood Ratio Test (LRT) can be used to derive a composition function that minimizes the average cost (Gu et al., 2008; Hoballah and Varshney, 1989). For this, Gu et al. (2008) defined the overall system output to be an alarm if $l(\vec{A}) > \tau$ and no alarm if $l(\vec{A}) < \tau$ (Eq. (1), where $\vec{A}$ refers to the vector containing the binary output of each classifier). τ is a constant based on the base-rate and cost ratio, while $l(\vec{A})$ is the likelihood ratio of the event. This strategy results in a minimal average cost under the assumption that both, true positives and true negatives, have no cost. Moreover, they highlighted that based on Neyman-Pearson theory, τ can be used to "maximize[] the probability of detection for a given upper bound on the false alarm rate" if the operation cost and base-rate should be unknown. Finally, Gu et al. (2008) proposed the likelihood ratio $l(\vec{A})$ as measure of suspicion, which enables the ranking of alarms for the operator.

$$\begin{aligned} l(\vec{A}) = \frac{P(\vec{A}|I)}{P(\vec{A}|\neg I)} > \frac{c_\alpha P(\neg I)}{c_\beta P(I)} = \tau \quad &\Rightarrow \text{output:} A \\ l(\vec{A}) = \frac{P(\vec{A}|I)}{P(\vec{A}|\neg I)} < \frac{c_\alpha P(\neg I)}{c_\beta P(I)} = \tau \quad &\Rightarrow \text{output:} \neg A \end{aligned} \tag{1}$$

Cost analyses are a valuable tool for specific deployments since they enable the comparison of different detection systems and provide practical operation point recommendations (Milenkoski et al., 2015). However, the results of any cost analyses follow from specifying the cost ratio, which often involves subjective estimations. Therefore, Gu et al. (2006b) introduced the Information-theoretic Framework for a more objective analysis approach, which will be described next.

2.2. Information-theoretic framework

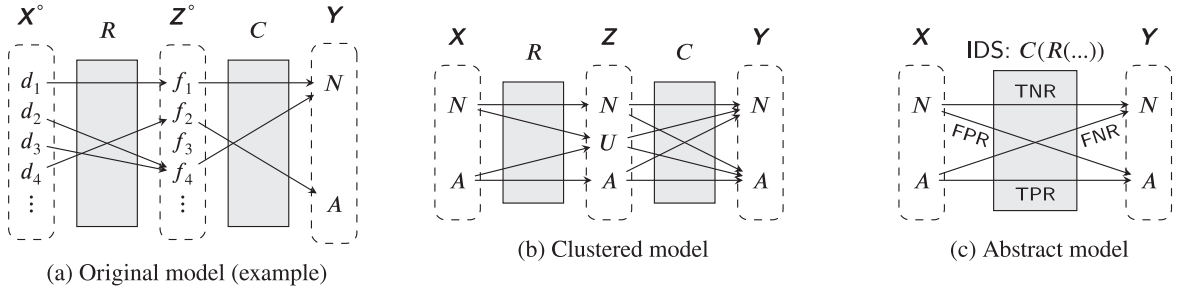
Gu et al. (2006a) first introduced the abstract Intrusion Detection System (IDS) model with the Intrusion Detection Capability and then extended it to the information-theoretic framework (Gu et al., 2006b). It aims to be a practical theory for an objective and data-driven analysis, which captures the most important aspects such as the TPR, FPR, PPV, NPV and base-rate to complement existing evaluation metrics.

Gu et al. (2006a,b) modeled the operation of a NIDS by three random variables X, Y, Z which generate data streams as shown in Fig. 1a. X represents the true state of the input data stream ($D = (D_1, D_2, \ldots)$), such that an oracle NIDS would assign $X_i = O_{\mathbf{NIDS}}(D_i)$. Z is the state of the intermediate feature representation ($Z_i = L_R(R(D_i))$, where R is the representation algorithm and L_R the feature representation labeling. The variable Y is the outcome of the classification algorithm C ($Y_i = C(R(D_i))$). This leads to the Markov chain $X \to Z \to Y$. The possible states of the random variables X and Y are $\{N, A\}$, where N represents normal and A anomalous samples. The possible feature representation states are $\{N, U, A\}$. U represents undistinguishability for samples of different classes, that are being mapped to the same feature vector. The representation labeling L_R shall therefore only give N and A to feature vectors (F) which can come from and only from one of the classes, while the label U shall be given if the feature vector could be from either a normal or anomalous class. Gu et al. (2006b, p. 534) formalized this using the following notation:

$$\begin{aligned} L_R(F_i) = N \Leftrightarrow \quad & \forall D_j, R(D_j) = F_i, O_{\text{NIDS}}(D_j) = N \\ L_R(F_i) = A \Leftrightarrow \quad & \forall D_j, R(D_j) = F_i, O_{\text{NIDS}}(D_j) = A \\ L_R(F_i) = U \Leftrightarrow \quad & \exists D_1 \neq D_2, R(D_1) = F_i, R(D_2) = F_i, \\ & O_{\text{NIDS}}(D_1) = N, O_{\text{NIDS}}(D_2) = A \end{aligned} \tag{2}$$

Information-theoretic measures can be defined based on this formalization for analyzing NIDSs. The Intrusion Detection Capability (C_{ID}) has been introduced as the normalized mutual information $C_{ID} = I(X;Y)/H(X)$ to measure the uncertainty reduction about the class of a sample from knowing the detection outcome, or how much ground truth information the NIDS can recover (Gu et al., 2006a). This analysis is based on the "abstract model" (Fig. 1c), which specifies a NIDS as tuple of its properties (bases-rate, FNR, FPR). Gu et al. (2006a,b) demonstrated that C_{ID} has a higher sensitivity to the relevant ranges of base-rates, FPR and TPR compared to the PPV, NPV and the probability of error (Gu et al., 2006a; 2006b). This makes C_{ID} suitable for finding an optimal operation point in objective comparisons by finding the point of the ROC curve which maximizes the intrusion detection capability.

Cardenas et al. (2006) related the intrusion detection capability C_{ID} back to the expected cost problem for finding the ideal operation point. They demonstrate that the operation point optimization based on the expected cost can be expressed by Eq. (3), where



(a) Original model (example) (b) Clustered model (c) Abstract model

Fig. 1. The Information-Theoretic Framework (adapted from Gu et al., 2006b, p. 532) represents the sample type (X), the feature representations (Z) and the detection outcome (Y) as random variables with the states normal/benign (N), anomalous/attack (A) and undistinguishable (U).

$c(X, Y)$ is the cost of each position in the contingency table.

$$r^* = \min_{(\text{TPR},\text{FPR})\in\text{ROC}} \mathbf{E}[c(X, Y)] \tag{3}$$

Equally, the operation point optimization by C_{ID} can be expressed by Eq. (4), as an instance of the expected cost problem with $c(X, Y) = -\log P[X|Y]$ (Cardenas et al., 2006).

$$\begin{aligned}(\text{TPR}^*, \text{FPR}^*) &= \arg\max_{(\text{TPR},\text{FPR})\in\text{ROC}} \tfrac{I(X;Y)}{H(X)} \\ &= \arg\min_{(\text{TPR},\text{FPR})\in\text{ROC}} \mathbf{E}[-\log P[X|Y]]\end{aligned} \tag{4}$$

Gu et al. (2006b) extended the analysis approach to the Information-Theoretic Framework for intrusion detection with the "clustered model" (Fig. 1b). The *feature representation capability* $C_R = I(X;Z)/H(X)$ can thus be defined as normalized mutual information between X and Z. Additionally, the *classification information loss* $L_C = I(X;Z|Y)/H(X)$ can be specified such that $C_{ID} = C_R - L_C$. This directly shows that the feature representation capability of a detection system leads to an upper bound on its detection capability ($1 \geq C_R \geq C_{ID} \geq 0$ and $1 \geq C_R \geq L_C \geq 0$) and enables a fine grained analysis or comparison of different NIDS components.

In 2013, Meng and Kwok (2013) applied the ITF to false-alarm filters in composed detection systems and proposed the *False Alarm Reduction Capability* $RC_{FA} = I(X_{\text{FA}}; Y_{\text{FA}})/H(X_{\text{FA}})$ as normalized mutual information between the filter input (X_{FA}) and output (Y_{FA}), equivalent to the intrusion detection capability C_{ID}.

By doing so, RC_{FA} is based on applying the ITF to false alarm filters as individual component rather than incorporating the false alarm filter into the ITF. The main drawback is, that RC_{FA} does not directly show how much the overall C_{ID} of the detection system improves by the additional component and also does not highlight the dependency between the different classifiers. These issues will be addressed further with an alternative approach to composed systems in Section 3, by extending the ITF of Gu et al. (2006b) rather than applying it to additional components.

2.3. Evaluating adversarial robustness

When deploying a detection system for security applications, the adaptive nature of adversaries or the limitations of the system evaluation should be considered. This is important since a detection system is of limited use if it can easily be evaded or does not perform as suggested. Robustness analyses have been approached differently in the literature depending on the research community (Cardenas et al., 2006; Puketza et al., 1996; Sheatsley et al., 2020). Approaches refer to the *sensitivity analysis*, which focuses on the impact of differing assumptions on the evaluation results, the *robustness of the classification method*, where perturbations of a sample without functional impact should not change the classification output (e.g. problem of defining a robust signature) or the *robustness of the implementation*, where the selected algorithms and resource constrains may impact the detection capabilities (e.g. problem of building a robust implementation of signature matching). This work focuses on the robustness of the classification method, but not the robustness of a specific implementation or the impact of resource constraints. Also notice the difference of the *robustness of the classification method* and *novelty detection* in this context. For this work, we define an attack as novel or previously unseen, if it can not be derived by the considered set of perturbation operations of the threat model and previous samples. Novelty detection will not be considered further, since it would require a different evaluation approach.

Most literature on NIDS focuses the robustness analysis on identifying the possible impact of evaluation limitations, differing assumptions and stress testing on the received results (Cardenas et al., 2006; Gu et al., 2006b; Puketza et al., 1996). One example for stress testing are *algorithmic attacks* which exploit the difference between the average and worst-case time complexity in Skip-Algorithms to cause packet drops from overloading (Zhang et al., 2013). However, since the stress testing based on a specific implementation, in this case for signature matching, is outside the scope of this work, it will not be considered further.

The issue of differing assumptions has been highlighted by Gu et al. (2006b) on the example of the base-rate. They noted that the ideal operation point depends on the base-rate, which is under adversarial control. The proposed solution was to dynamically adjust the operation point based on a base-rate estimation from the alarm rate.

They also proposed to analyze the impact of evaluation limitations by considering uncertainty ranges for parameters like base-rate, TPR or FPR and find the ideal worst-case performance using the information-theoretic framework. Similarly, Cardenas et al. (2006) proposed a $(\delta_p, \delta_\alpha, \delta_\beta)$-intruder for the robustness analysis of a specific system on the IDOC. A $(\delta_p, \delta_\alpha, \delta_\beta)$-intruder can change its base-rate within the bounds of $\delta_p = [p - \delta_{p_l}, p + \delta_{p_u}]$ and cause misclassified samples such that the performance is reduced to $\text{TPR}' = \text{TPR} \cdot (1 - \delta_\beta)$ and $\text{FPR}' = \delta_\alpha + \text{FPR} \cdot (1 - \delta_\alpha)$. This already shows how the parameter ranges of Gu et al. (2006b) are equivalent to a set of $(\delta_p, \delta_\alpha, \delta_\beta)$-intruders by Cardenas et al. (2006) and how the maximal parameter uncertainty leads to a worst-case intruder.

However, both of these approaches are based on assumptions about the uncertainty in the evaluation and confidence in the specific detection system. Therefore, the analysis is subjective and might bias the results based on prior believes. This issue could be solved by evaluating specific attacks, like it is commonly done in the area of adversarial robustness.

Recent work focused on evaluating the adversarial robustness of NIDS from a machine learning perspective on traffic shaping. Adversarial examples can be generated by solving the following optimization problem (Eq. (5)). This aims on finding a sample $\vec{x^*}$ with minimal distance to the original sample $\vec{x}$, while causing a misclassification ($T(\vec{x^*}) \neq T(\vec{x})$) on the target classifier (T) (Papernot et al., 2017). This solution can be approximated by different methods

(typically gradient-based), while the distance $\delta_{\vec{x}}$ is constrained with an ℓ_p norm (Carlini et al., 2019). However, there seems no consent which ℓ_p constraint would be suitable for the application. Sheatsley et al. (2020) argued for an ℓ_0 norm since only limited features in a packet sequence can be perturbed, while Hartl et al. (2020) considers ℓ_1 and ℓ_∞ norms as practically relevant. This issue will be addressed further by relating the distance to a practical threat model in Section 3.3.

$$\vec{x}^* = \vec{x} + \delta_{\vec{x}} = \vec{x} + \arg\min\{\vec{z} : T(\vec{x}+\vec{z}) \neq T(\vec{x})\} \tag{5}$$

According to Carlini et al. (2019), adversarial robustness should be evaluated using the strongest known attack with adaptation to the specific countermeasures. Several methods have been proposed which adopt the generation of adversarial examples to the constraints of network traffic. Hartl et al. (2020) generated adversarial examples by restricting the allowed perturbations to the inter-arrival time and packet length within logical constraints of the traffic under control. Hashemi et al. (2019) and Sheatsley et al. (2020) additionally considered the relation of primary and dependent features. These dependencies between features were either explicitly specified (Hashemi et al., 2019) or heuristically derived from a data set into firsts-order logic (Sheatsley et al., 2020).

The robustness is then typically evaluated based on the trade-off between the achievable worst-case performance loss and the required perturbation budget (Carlini et al., 2019). Hashemi et al. (2019) evaluated the TPR reduction at a specified FPR. Similarly, Hartl et al. (2020) evaluated the TPR reduction and trade-off between the success-rate and required perturbation distance. Additionally, they introduced the Adversarial Robustness Score (ARS) as the average distance of the 50% adversarial examples with minimal distance. This causes the ARS to become infinite, if the attack fails for more than half of all samples. Sheatsley et al. (2020) also evaluated the attack success rate depending on the allowed perturbation distance and used this to additionally study the inter-/intra-transferability of adversarial examples between classifiers.

Adversarial examples have also been applied in the context of ensemble classifiers. In the image processing domain, Tramér et al., 2020 highlighted that augmenting the training data with adversarial examples (adversarial training) to increase the model robustness can create misleading results, since the model might remain vulnerable to other attack methods. They aimed on reducing this issue by introducing *Ensemble Adversarial Training*, which additionally augments the training data for a model with adversarial examples that were generated for individual classifiers within a set of static pre-trained models (ensemble). Hang et al. (2020) proposed specific black-box attack methods, which generate ensemble substitute classifiers for the target to generate adversarial examples based on a boosting structure (selective cascade ensemble strategy) or bagging structure (stack parallel ensemble strategy).

Discussion: One resulting question is, if it is possible to specify an equivalent $(\delta_p, \delta_\alpha, \delta_\beta)$-intruder for an adversarial example attack method. While both approaches depend on the detection system, the approach by Gu et al. (2006b) and Cardenas et al. (2006) were based on assuming a general undetectability, while adversarial examples typically reduce the original detection score. This implies that the performance of a $(\delta_p, \delta_\alpha, \delta_\beta)$-intruder is independent of the used operation point, while the performance evaluation on adversarial examples is operation point dependent. Therefore, it would be possible to specify an equivalent $(\delta_p, \delta_\alpha, \delta_\beta)$-intruder with additional operation point dependence for an adversarial attack on a detection system.

While the evaluation with adversarial examples results in an objective analysis by specifying a threat model and attack method, the impact on parameters of interest at low base-rates have not been investigated further like by the NIDS literature. The robustness evaluations were also limited to the classification of attack samples, while studying the base-rate fallacy indicates that a detection system would become equally unusable if an arbitrary amount of false alarms could be caused. Therefore it would be important to combine the different evaluation approaches. This will be discussed further in Section 3.3.

3. Information-theoretic framework of intrusion detection for composed systems and robustness analyses

To avoid the limitations from combining the Intrusion Detection Capability (Gu et al., 2006b) and False Alarm Reduction Capability (Meng and Kwok, 2013) for analyzing composed systems, their combination is briefly discussed to identify the cause of resulting issues. This leads to the introduction of an alternative representation for composed NIDSs, which is used to extend the information-theoretic framework. Afterwards, the resulting opportunities for studying the system robustness are highlighted and the properties of the evaluation metrics are discussed to further increase the interpretability of results.

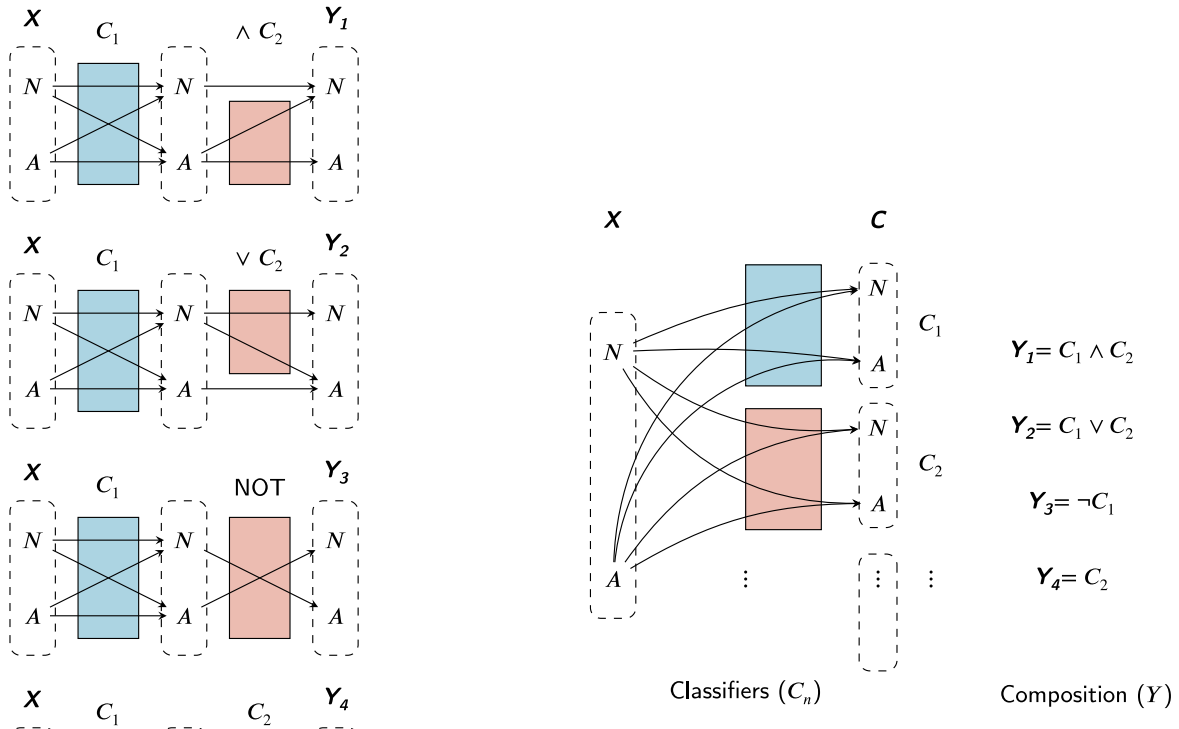
3.1. Representing composed classifiers

Meng and Kwok (2013) applied the information-theoretic framework for the evaluation of false alarm filters. The proposed *False Alarm Reduction Capability* can be expressed in the context of the full detection system as $RC_{FA} = I(X; Y_{FA}|Y_C = A)/H(X|Y_C = A)$. Here, X is the input state, Y_C the detection system outcome and Y_{FA} the false alarm filter outcome. However, due to the condition of $Y_C = A$, RC_{FA} does not directly relate to its impact of the overall C_{ID} and the resulting RC_{FA} score has a dependency on both, the primary detection system Y_C and the specific arrangement. Therefore it would be desirable to extend the ITF for composed NIDSs.

The key challenge by incorporating chained detecting systems into the ITF is that a processing sequence $X \to Z \to Y_C \to Y_{FA}$ is not a valid Markov chain. The false alarm filter Y_{FA} and the feature representation Z are not conditionally independent given Y_C, which leads to the contradiction $I(X|Y_{FA}) \not\leq I(X|Y_C)$ that violates the data processing lemma.

This issue can be addressed by converting the sequential representation of a composed NIDSs into an equivalent parallel representation as shown in Fig. 2. The sequential representation can refer for example to the system architecture, where a firewall filters which samples reach following detection methods. The parallel representation on the other hand views the same system as an equivalent ensemble of classifiers. Independent of the arrangement or representation, each classifier operates on its own feature representation.

The previous example of a false alarm filter (Y_1 in Fig. 2 a) which determines the transitions of alarms from the prior detection method (C_1) to the system output (Y_1), is equivalent to applying an $\wedge$-gate as composition function to the output of both classifiers (Y_1 in Fig. 2 b). Similarly could be viewed for example a blocklist (C_1) before a detection system (C_2) as applying an $\vee$-gate as composition function to the output of both classifiers (Y_2 in Fig. 2 a/2 b). These examples highlight how any arrangement of classifiers can be expressed by a Boolean function as shown in Fig. 2. This concept can be generalized to any number of composed classifiers. The system is represented as a layer of parallel classifiers generating the state $C = (C_1, C_2, \ldots)$, which is then aggregated by a single composition function that captures the arrangement of the individual classifiers as Boolean equation. This enables a direct extension of the ITF for composed classifiers, since it becomes possible to define a valid Markov chain. This will be discussed further in Section 3.2.



(a) Sequential representation (b) Equivalent parallel representation

Fig. 2. Composed system representations: Any system of chained classifiers can be represented as an equivalent ensemble with the respective composition function.

3.2. Information-theoretic framework for composed classifiers

The information-theoretic framework can be extended for composed classifiers as shown in Fig. 3, which we will simply refer to as *composed ITF*. It considers a random variable X, with normal samples (N) and anomalous samples (A) from their respective sources (S_N/S_A). The samples of X are being processed by n *feature representation algorithms*, one for each classifier, to the new variable $Z = (Z_1, ..., Z_n)$, which considers the states $Z_i = \{A, U_1, ..., U_x, N\}$ of a feature vector. Each label U_j represents a distinct set of undistinguishable samples in this feature representation. This definition for undistinguishability had to be changed from Eq. (2) to satisfy the required properties of the Markov chain model, as shown in Appendix 1. This leads to the adaptation of the notation by Gu et al. (2006b, p. 534) which is shown in Eq. (6). Rather than one label for undistinguishability (U), it allows for multiple U_x and beyond expecting all feature vectors of the same U_x to be equal and contain different classes, it requires the feature vectors of different U_x to be different ($\forall k \neq x, \forall D_i \in U_x, \forall D_j \in U_k, R(D_i) \neq R(D_j)$).

$$\begin{aligned} L_R(F_i) = N \Leftrightarrow \quad & \forall D_j, R(D_j) = F_i, O_{\text{NIDS}}(D_j) = N \\ L_R(F_i) = A \Leftrightarrow \quad & \forall D_j, R(D_j) = F_i, O_{\text{NIDS}}(D_j) = A \\ L_R(F_i) = U_x \Leftrightarrow \quad & \exists (D_1, D_2), R(D_1) = F_i, R(D_2) = F_i, \\ & O_{\text{NIDS}}(D_1) = N, O_{\text{NIDS}}(D_2) = A, \\ & \forall k \neq x, \forall D_j \in U_k, R(D_j) \neq F_i \end{aligned} \tag{6}$$

The samples of Z are being processed by a set of m *classifiers* to the new variable $C = (C_1, ..., C_m)$, which contains the outcome of each classifier $C_x = \{A, N\}$. The specific *arrangement or composition* of the classifiers can then be represented with a boolean function as demonstrated in Section 3.1, which leads to the final detection outcome Y. This approach models a composed NIDS as Markov chain $X \rightarrow Z \rightarrow C \rightarrow Y$.

Based on this, we can adopt and extend the definitions of the clustered model from Gu et al. (2006b). The Definitions 1 and 3 are identical to the definition of Gu et al. (2006b), while Definition 5 is adjusted from Gu et al. (2006b) to the new Markov chain.

Definition 1. The *intrusion detection capability* $C_{ID} = \frac{I(X;Y)}{H(X)}$ is the normalized mutual information between the input (X) and output (Y).

Definition 2. The *classification capability* $C_C = \frac{I(X;C)}{H(X)}$ as normalized mutual information between the input (X) and all classifier outputs (C).

Definition 3. The *feature representation capability* $C_R = \frac{I(X;Z)}{H(X)}$ is the normalized mutual information between the input (X) and the feature representations (Z).

Definition 4. The *composition information loss* $L_Y = \frac{I(X;C|Y)}{H(X)}$ as normalized conditional mutual information between the input (X) and all classifier outputs (C) given the final detection outcome (Y).

Definition 5. The *classification information loss* $L_C = \frac{I(X;Z|C)}{H(X)}$ as normalized conditional mutual information between the input (X) and the feature representations (Z) given all classifier outputs (C).

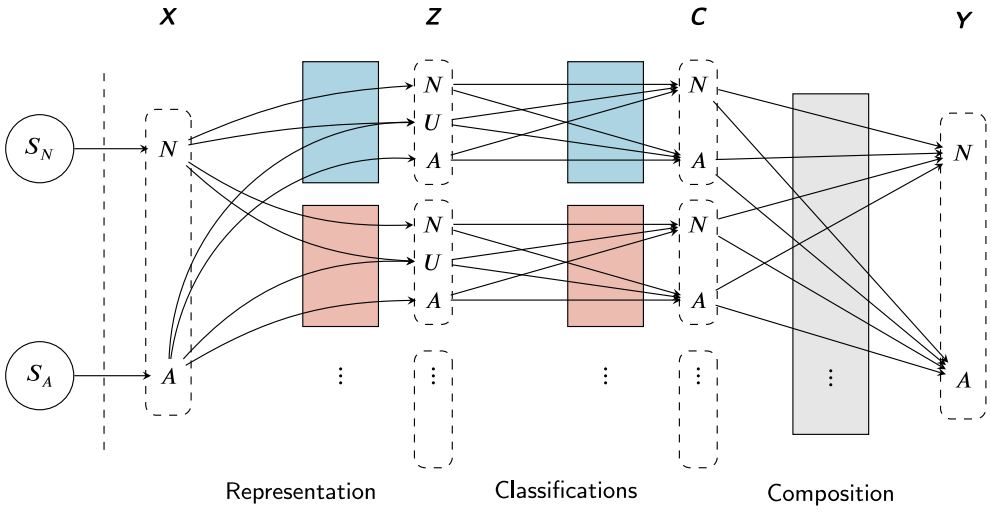


Fig. 3. Information-theoretic framework for composed classifiers (only one U shown per feature representation).

Definition 6. The *classifier gain* $G_{C(1..n,n+1)} = \frac{I(X;C_{n+1}|C_{1..n})}{H(X)}$ as normalized conditional mutual information between the input (X) and classifier C_{n+1}, if the outcome of the classifiers $C_{1..n}$ are already known.

Definition 7. The *representation gain* $G_{R(1..n,n+1)} = \frac{I(X;Z_{n+1}|Z_{1..n})}{H(X)}$ as normalized conditional mutual information between the input (X) and representation Z_{n+1}, if the representations $Z_{1..n}$ are already known.

Since the model is a Markov chain, it directly follows from the Data Processing Lemma that: $1 \geq C_R \geq C_C \geq C_{ID} \geq 0$ and $C_{R_i} \geq C_{C_i}$. This demonstrates how the feature representation (C_R) presents an upper bound for the intrusion detection capability (C_{ID}) and that no composition of the given classifiers could achieve a better performance than their total classification capability (C_C).

Moreover, the definitions lead to additive properties to enable an attribution of the overall system performance to its individual components as shown in Appendix 2. The intrusion detection capability results from the feature representation capability and losses from the classifiers and their composition $C_{ID} = C_R - L_C - L_Y = C_C - L_Y$. Adding an additional feature representation or classifier increases the respective capability with its gain, such that $C_{C_{(1,2)}} = C_{C_1} + G_{C(1,2)}$ and $C_{R_{(1,2)}} = C_{R_1} + G_{R(1,2)}$ where C_1 and C_2 are two classifiers $C = (C_1, C_2)$ and Z_1 and Z_2 are two feature representations $Z = (Z_1, Z_2)$.

The relations between the used information theoretic metrics are visualized for a composed system with two classifiers in Fig. 4. Notice that both classifiers could themselves be composed of an arbitrary number of classifiers.

Bringing this back to the example of evaluating a false alarm filter, it can be seen that a second classifier can not improve the overall system performance by more than $G_{C(1,2)}$ ($C_{ID} \leq C_{ID_1} + G_{C(1,2)}$). This highlights that $G_{C(1,2)}$ or $G_{C(1,2)} - L_Y$ could be a suitable metrics by evaluating and comparing possible false alarm filters that complement an individual detection system. Additionally should be highlighted that the conditional mutual information $I(X; C_2|C_1)$ of a second classifier C_2 can be both, bigger or smaller, than its mutual information $I(X; C_2)$.[4] Therefore, the classification gain $G_{C(1,2)}$ of classifier C_2 can also be both, bigger or smaller, than its classification capability (C_{C_2}) and individual intrusion detection capability (C_{ID_2}).

By evaluating specific components, it is important to view the achieved performance in relation to their limits. A composition loss of $L_Y = 0$ is not always possible due to the reduction of states between the random variables. However the achievable minimum can often easily be found due to a small number of total states.

Discussion: Besides the redefinition of the label U, the extended definitions become identical to the work of Gu et al. (2006b), if there exists only one classifier ($Y = C$) which leads to $L_Y = 0$. Like Gu et al. (2006b) also noted, the optimal operation point depends on the base-rate, which is under adversarial control. The proposed solution (Gu et al., 2006b) was to dynamically adjust the operation point based on a base-rate estimation from the alarm rate. Since an adjustment on the operation point changes the TPR and FPR of the classifier, it may also change how to ideally compose the different classifiers. Therefore, the thresholds and the composition should be adjusted dynamically to the base-rate, but this needs to be done with care since an operation point adjustment based on alarms could be exploited by adversaries.

Finally, the robustness approach of Gu et al. (2006b) and Cardenas et al. (2006) can also be adopted equally, by analyzing ranges for the base-rate, TPR and FPR or specifying a $(\delta_p, \delta_\alpha, \delta_\beta)$-intruder and selecting the best worst-case performance. However, the information-theoretic framework enables a more detailed analysis as it will be discussed in Section 3.3.

3.3. Information-theoretic framework for adversarial robustness

The information-theoretic framework can be used to perform fine grained robustness analyses. This leads to the advantage that the adversarial robustness can be studied by using metrics out of the intrusion detection area and can be applied to both individual and composed classifiers. Studying composed classifiers might be additionally interesting, since the robustness of the composed system could rely heavily on the performance of individual classifiers. Before highlighting which dynamics between classifiers can be identified with the framework, the term ideal robustness will be defined further and its relation to the pre-processing in NIDSs will be discussed. However, the analysis first requires to define the term "robustness" in respect to a threat model.

A threat model shall contain a set of *perturbation operations* which can be applied to a sample for causing a misclassifications without affecting the sample's functionality. In addition, *practical constraints* can be specified on the adversarial capabilities, like for

[4] An example of synergy and redundancy can be found in Bossomaier et al. (2016, p. 43).

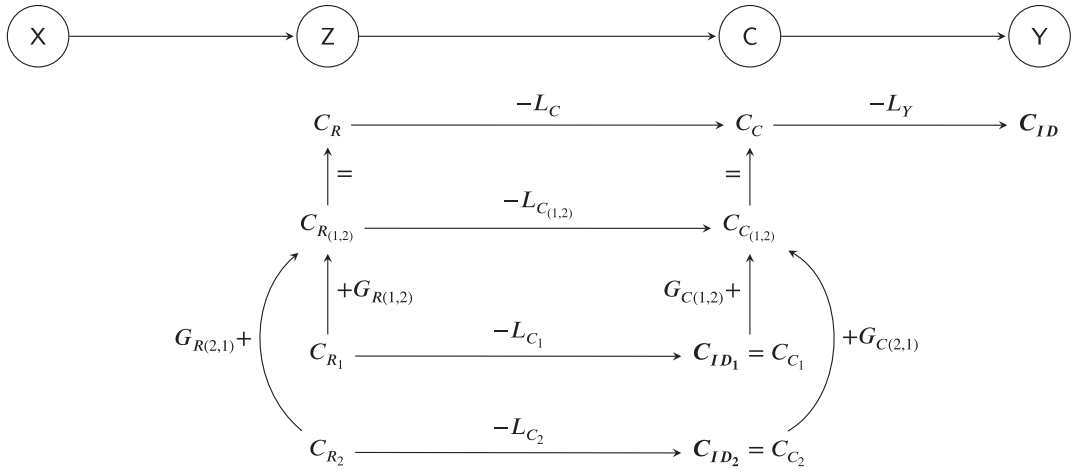


Fig. 4. Relation between the information theoretic metrics: The system capabilities ($C_{R_{(1,2)}}$, $C_{C_{(1,2)}}$) can be broken down to the capabilities and gains of its components.

example a maximal bandwidth, maximal time or which fraction of samples at the input are under adversarial control and which *system knowledge* is given.

The separation of perturbation operations and the attack budget enables the distinction of an adversarial sample's *distance* and *cost* for the adversary. The distance between two samples can be expressed by using an ℓ_p-norm or edit distance and is typically constrained by generating adversarial examples. For practical relevance, it would however be of greater interest to instead constrain the *cost* for the adversary to find samples within a given attack budget of the threat model. For the threat model example above, this could be the maximal bandwidth and flow duration.[5]

Finally, the information-theoretic framework could consider benign or adversarial sources as shown in Fig. 5. While a benign source would directly generate benign samples, an adversarial source can generate both, benign and attack samples, and has the additional capability of perturbing the samples without affecting their functionality by using the operations of the threat model. While it is often only considered how adversarial perturbations could hide an attack, the issue of the base-rate fallacy indicates that the possibility to increase the FPR would equally question the usability of the system.

The framework leads to the definition that a classifier is *ideally robust* against a threat model, if no allowed input perturbations (P) can lead to a misclassification. Notice that this definition of robustness does *not* require the classifier to correctly identify new samples from the same or other attack classes. Instead it requires a consistent classification output for all allowed perturbations of the same sample. For a meaningful robustness analysis, this would require a careful study of the possible perturbation operations for each network layer up to the specific application and attack.

Ideally robust worst-case performance: The first question which can be studied is how well an ideally robust classifier can maximally perform *on the given data set*, if all attack samples and a fraction q of benign samples came from an adversarial source. To answer this question, we need to redefine when two samples are *undistinguishable* for an ideally robust classifier.

Definition 8. A sample x of source S_x is undistinguishable between the sources S_x and S_y in an adversarial environment, if its feature representation is reachable from some sample of S_y or if the sample can reach the feature representation of some sample of S_y through the perturbation operations (P) of the threat model.

$$\begin{aligned} L_R(F_i) = U_x \Leftrightarrow \quad & \exists (D_1, D_2), R(P(D_1)) = F_i, R(P(D_2)) = F_i, \\ & O_{\text{NIDS}}(D_1) = N, O_{\text{NIDS}}(D_2) = A, \\ & \forall k \neq x, \forall D_j \in U_k, R(P(D_j)) \neq F_i \end{aligned} \tag{7}$$

By using this definition of undistinguishability, the resulting feature representation capability C'_R will be an upper bound on the performance of any ideally robust classifier given this particular data set and threat model. This also implies that there always exists an adversarial attack which could at least decrease the performance by $\max(C_{ID} - C'_R, 0)$. Since this demonstrates that a classifier can not be ideally robust if $C_{ID} > C'_R$, a low C'_R directly indicates that the used feature representation is unsuitable for a robust classification; independent of the used detection methods.

Achieving ideal robustness through pre-processing: The previous analysis already indicates, that any non-robust classifier can be converted into an ideal robust classifier through the use of data pre-processing and normalization. Some typical examples of practical systems would be the case conversion during signature matching or a target-based packet reassembly. The aim of this pre-processing is to map every possibly equivalent sample under the given set of operations to an identical feature representation ($\forall D_j, R(P(D_j)) = R(D_j)$). Another equivalent formulation would be that the sample x shall receive an identical feature representation to every possible sample that would be reachable with the perturbation operations of the threat model. This would automatically lead to $C_{ID} \leq C'_R$. Moreover, it guarantees ideal robustness for the classification algorithm against the perturbations of the threat model since all derived samples are undistinguishable in their feature representation. This highlights how pre-processors and normalizations are useful tools for improving system robustness against a threat model.

Evaluating the robustness of detection systems: While the previous analyses were purely based on the data set and threat model to provide bounds on the achievable performance, the main aspect of interest is of course the robustness of the detection method. This aims on finding sample perturbations which would lead to a misclassification by the system under test. One key difference to the analysis before is, that the analyzed samples could be outside of the given data set but possible under the threat model. Since the feature space is too large for searching, gradient-based methods are typically used for approximating a possible sample. This requires the use of a specific attack method and it should be remembered that the following results only provide a lower bound on the achievable performance degradation as stronger attack meth-

[5] This example is similar to the approach of Hartl et al. (2020), who constraint the ℓ_1 distance by perturbing the packet timing and packet size increase.

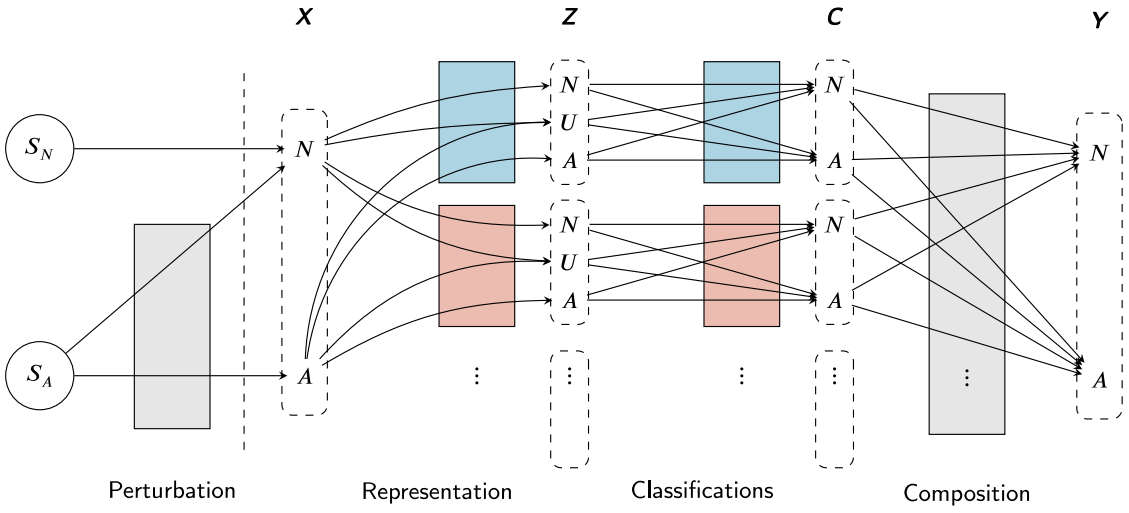


Fig. 5. Information-theoretic framework for the robustness analyses (only one U shown per feature representation).

ods may exist. To evaluate the overall system robustness, adversarial examples can be generated for all classifiers simultaneously. However, it is also possible to perform further analyses which can typically be found in the area of adversarial robustness, such as the dependence on constraints or transferability to other system components.

The key advantage of the information-theoretic framework is that it provides detailed information on how the total performance degradation of the attack is split over all system components, such as the feature representation, individual classifiers, dependencies between classifiers or the used composition. This makes it a useful tool for designing robust systems by identifying and attributing possible weaknesses to the individual components of the NIDS. Similarly can the performance degradation of an adversarial attack ($C_{ID} - C'_{ID}$) be separated into its components which are detection method independent ($C_R - C'_R$) and detection method dependent ($L'_C - L_C$ and $L'_Z - L_Z$).

How the redundancies between classifiers are utilized during the attack or how synergies and operation point changes amplify the caused damage can be of additional interest in the context of composed systems. For example, a classification gain $G_{C(1,2)} < C_{C_2}$ indicates that the classifier C_2 could provide redundancy for the classifier C_1 during an attack, if they were sufficiently independent. In this case, an increase of the classification loss L_{C_1} can lead to an increase in the information gain $G_{C(1,2)}$. This would be a desirable behaviour and can be evaluated with the analysis on an adversarial data set. If the classification redundancy could be utilized during the attack, then either the classifiers were sufficiently independent or the transferability of the attack method was insufficient - depending on the perspective.

On the other hand, an information gain $G_{C(1,2)} > C_{C_2}$ indicates a synergy-based performance dependency of classifier C_2 on C_1. In this case, evading only classifier C_1 can additionally decrease the information gain of classifier C_2. This highlights that the classification capability C_{C_1} does not provide an upper bound on the impact of its evasion on the overall decrease in classification performance ($C_C - C'_C$) which can be up to $C_{C_1} + G_{C(1,2)} - C_{C_2}$.

The analysis on an adversarial data set can also be used to identify which operation points would maintain the highest C_{ID} while being under attack and how the ideal composition changes. While this design approach may limit the performance during normal operation, it could be used to increase the system robustness. Similarly, the impact of a different data pre-processing and normalization can be quantified and compared by evaluating how much they improve the robustness of the classification algorithms C'_C or C'_{C_x}.

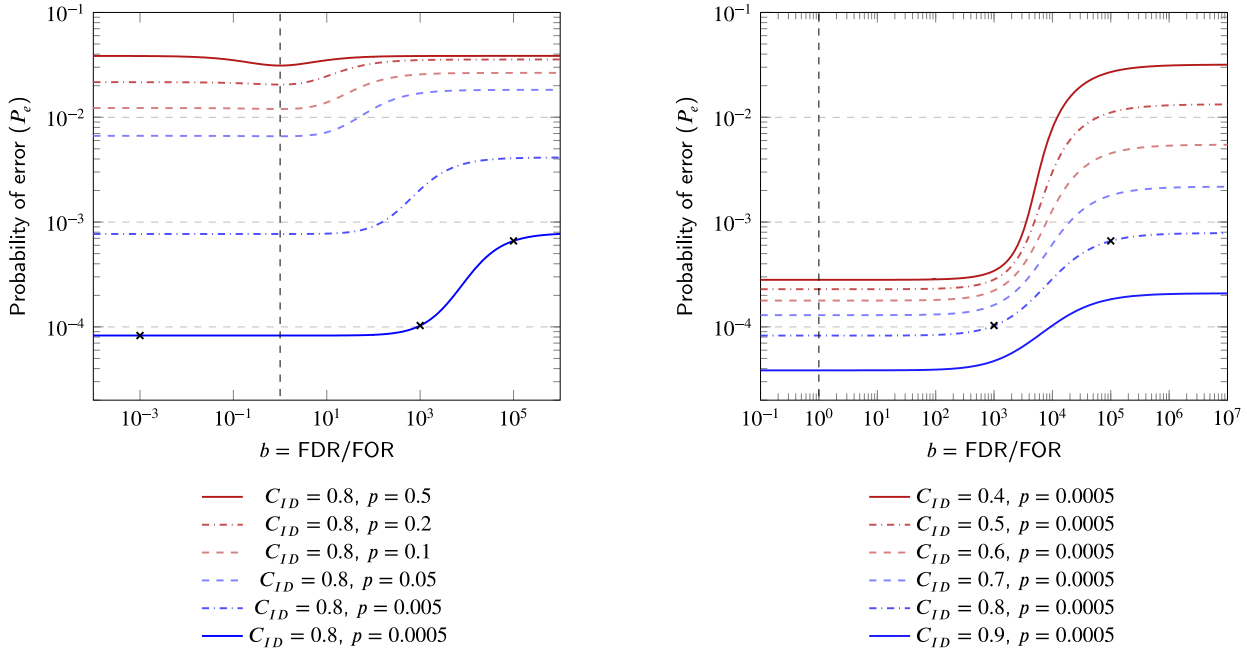
Evaluating the risk of adversarial examples: Not all adversarial samples present a similar risk in practice, since they could be destroyed by a given channel with, for example, timing jitters or packet errors. Therefore, it could be performed an analysis with a specified channel model, that additionally disturbs the samples like a communication channel without control of the adversary to evaluate the risk from an attack method.

3.4. Properties of the intrusion-detection capability

Gu et al. (2006b) highlighted that the normalized mutual information measures the *uncertainty* reduction about the source of a sample after its classification outcome is known. Cardenas et al. (2006) noted additionally that this operation point selection is an instance of the expected cost analysis with $c(X, Y) = -\log P[X|Y]$. Gu et al. (2006a,b) also demonstrated that the C_{ID} has a higher sensitivity to the relevant ranges of base-rates, FPR and TPR compared to the PPV, NPV and probability of error P_e. The following section will highlight another property of C_{ID} and its relationship to P_e.

Towards a skill metric: In the context of evaluating forecast systems for rare events, an area with similar challenges by having high impact events at a strong class imbalance, *equitability* is a desired property that might also be desirable for the evaluation of NIDSs. Gandin and Murphy (1992) defined a measure as equitable if any random or constant classifier receives an identical score and combines the elements of the contingency table by a linear weighted sum (Hogan et al., 2010). Hogan et al. (2010) noted however that the linearity requirement can be omitted, which enables the definition of measures that avoid vanishing results at decreasing base-rates. An example for a measure that assigns different scores to random or constant classifiers would be the expected cost. At a constant cost ratio c and a base-rate p, it would assign $c_{\text{exp}} = cp$ to a system that never returns an alarm and $c_{\text{exp}} = 1 - p$ to a system that always returns an alarm. In this case, there may exist systems with skill that result in a higher cost than others without skill. While this directly relates to their practical value, it can be undesirable in objective comparisons between systems. This issue equally applies to the probability of error P_e or using the FPR, TPR, PPV and NPV individually.

In the ITF, any random or constant classifier can be represented by an independent random variable as generator of C. In this case $C_C = \frac{I(X,C)}{H(X)} = \frac{(H(X)-H(X|C))}{H(X)} = 0$ leading to $C_{ID} = 0$, since the conditional entropy is $H(X|C) = H(X)$ as X and C are independent by the definition of a random or constant classifier. This highlights that



(a) Constant C_{ID} - dependence of P_e/b on the base-rate p

(b) Constant p - dependence of P_e/b on C_{ID}

Fig. 6. C_{ID} and its lower bound on the probability of error (× indicates the parameters of Table 1): P_e is minimized at $b = 1$, systems with higher P_e can sometimes achieve a lower C_{ID} and at low base-rates a higher b might allow higher P_e at the same C_{ID}.

the normalized mutual information results in a score of 0 for any random or constant classifier, while any system with skill would receive better results since a dependence between X and C would reduce the conditional entropy $H(X|C) \leq H(X)$.

Relation of C_{ID} ***and*** P_e: To further interpret the normalized mutual information, its relation to the probability of error $P_e = P(X \neq Y)$ can be highlighted. Assuming a constant base-rate for the evaluation, the intrusion detection capability only depends on the conditional entropy $H(X|Y)$ since the entropy of X only depends on the base-rate ($H(X) = -p\log_2(p) - (1-p)\log_2(1-p)$).

$$C_{ID} = \frac{I(X,Y)}{H(X)} = \frac{H(X) - H(X|Y)}{H(X)} \tag{8}$$

The conditional entropy can be related to the probability of error with Fano's inequality, where $H(E) = -P_e\log_2(P_e) - (1-P_e)\log_2(1-P_e)$ is the binary entropy of the error probability and N the number of possible states (Fano, 1961, p. 186).

$$H(X|Y) \leq H(E|Y) + P_e\log(N-1) \leq H(E) + P_e\log(N-1) \tag{9}$$

The case of a binary classification problem (benign/attack, $N = 2$) simplifies the equation to $H(X|Y) \leq H(E)$. The equality $H(X|Y) = H(E) + P_e\log(N-1)$ holds if the probability of an error given that $Y = y$ is equal for all y and if all remaining values $X = x$ are equally likely in the case of an error (Massey, 1998, p. 84) (Fano, 1961, Eq. 6.19). The second condition always holds for a binary classification problem since there is only one alternative state in case of an error. The first condition holds if $P(X = A|Y = N) = P(X = N|Y = A)$, which applies to systems where the False Omission Rate (FOR) equals the False Discovery Rate (FDR). This directly implies that C_{ID} can be associate with a minimal error of probability or in other words that the probability of error has a lower bound for all systems with identical C_{ID}.

Fig. 6 visualizes the relation of C_{ID} to the probability of error in the context of the base-rate fallacy. It presents the probability of error as a function of C_{ID}, the base-rate p and the ratio $b = \text{FDR/FOR}$ to represent the missclassification symmetry. This ratio enables that the minimum of the error probability can be found at $b = 1$ for any C_{ID} and p, which is highlighted by a vertical dashed line.

Fig. 6 a demonstrates the impact of the base-rate fallacy and plots the function of P_e for a constant C_{ID} at decreasing base-rates. It can be seen how the symmetry at an equal class balance ($p = 0.5$) breaks for decreasing base-rates and how a missclassification symmetry of FOR > FDR leads to diminishing possible increases in P_e at the same intrusion detection capability for lower base-rates. Since C_{ID} is a complementing evaluation metric, it can be found a further analysis of marked examples (× in Fig. 6) that would achieve $C_{ID} = 0.8$ at a base-rate of $p = 0.0005$ for a ratio b of 10^{-3}, 10^3 and 10^5 in Table 1. As expected, it can be seen that lower ratios of b relate to systems with higher PPV and that a high PPV at low base-rates requires diminishing FPRs. It can also be seen that the higher P_e at $b = 10^5$ is caused by a higher FPR, lowering the PPV and requiring a higher detection rate to maintain C_{ID}.

Fig. 6 b shows the probability of error for several C_{ID} at a low base-rate of $p = 0.0005$. Since the increase in P_e is diminishing for $b < 1$, the plot has instead been extended for higher ratios. This demonstrates how a reduced intrusion detection capability allows a higher probability of error. However, it can also be seen that a system with higher P_e may achieve the same or better uncertainty about the input than a system with lower P_e, if the ratio of FDR/FOR becomes sufficiently large. This directly implies that using Bayes decision rule to derive a composition function will lead to a minimal P_e, but might not achieve a maximal C_{ID}. An example scenario for this can be found in Appendix 3.

This highlights that C_{ID} can be a suitable measure for *detection skill*, as it avoids vanishing results at decreasing base-rates and assigns a minimal score to any random or constant classifier. Moreover, the relation of C_{ID} to P_e, TPR and PPV has been discussed to increase its interpretability as complementing evaluation metric and demonstrate their dependencies.

Relation of C_{ID} ***and cost analyses in ROC plots:*** Since the relation of C_{ID} and cost analyses has been discussed by Gu et al. (2006a) and Cardenas et al. (2006), the composed ITF

Table 1
Example systems with $C_{ID} = 0.8$ at $p = 0.0005$.

P_e	b	p	C_{ID}	TPR	FPR	PPV	NPV	FOR	FDR
8.27e-5	0.001	0.0005	0.80	0.835	3.45e-11	1 - 8.27e-8	1 - 8.27e-5	8.27e-5	8.27e-8
1.03e-4	1000	0.0005	0.80	0.859	3.25e-5	0.93	1 - 7.03e-5	7.03e-5	7.03e-2
6.61e-4	100,000	0.0005	0.80	0.989	6.55e-4	0.43	1 - 5.70e-6	5.70e-6	0.57

Table 2
System example with two classifiers.

C \ X	N	A	l(C)	τ
(N, N)	0.18	0.05	1.11	τ_3
(A, N)	0.5	0.02	0.16	τ_4
(N, A)	0.05	0.07	5.6	τ_1
(A, A)	0.07	0.06	3.43	τ_2

(a) Joint probabilities

C	f_1	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}	f_{14}	f_{15}	f_{16}
(N,N)	N	N	N	N	N	N	N	N	A	A	A	A	A	A	A	A
(N,A)	N	N	N	N	A	A	A	A	N	N	N	N	A	A	A	A
(A,N)	N	N	A	A	N	N	A	A	N	N	A	A	N	N	A	A
(A,A)	N	A	N	A	N	A	N	A	N	A	N	A	N	A	N	A
=	N	AND	$C_1 \wedge \overline{C_2}$	C_1	$\overline{C_1} \wedge C_2$	C_2	XOR	OR	$\neg f_8$	$\neg f_7$	$\neg f_6$	$\neg f_5$	$\neg f_4$	$\neg f_3$	$\neg f_2$	A

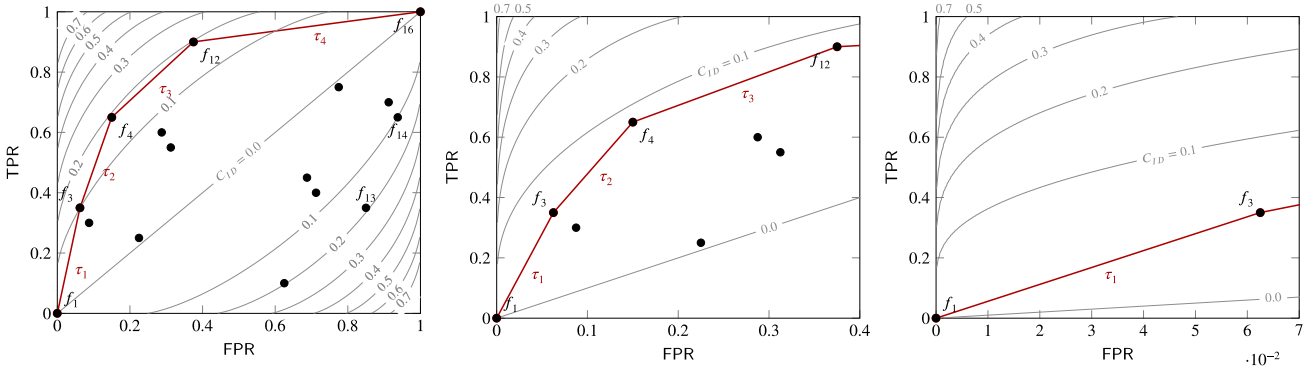
(b) Enumeration of the possible composition functions

will now be set into the context of the cost-based alert fusion from Gu et al. (2008) by visualizing them in ROC plots.

Even though Cardenas et al. (2006) highlighted that maximizing C_{ID} is an instance of the expected cost problem (see Section 2.2), the approach of Gu et al. (2008) with using the LRT (Hoballah and Varshney, 1989, Eq. 8) as optimal solution for deriving a composition function from a cost perspective, is not directly applicable to C_{ID} since it does not have a fixed cost (the cost $c(X, Y) = -\log P[X|Y]$ depends on the unknown composition function). Nevertheless, using the likelihood ratio to derive an ideal composition function from an information-theoretic perspective still provides further insights on their relation, as it will be intuitively explained with an example.

Consider a system of two classifiers which have the joint probabilities shown in Table 2a and all possible composition functions shown in Table 2b. The used base-rate for this example is $p = 0.2$ to simplify the visualization. Fig. 7a shows the corresponding ROC plot containing the result of all composition functions and isolines for systems of identical C_{ID} (gray).

As expected, it can be seen that the composition functions f_1 and f_{16} result in a constant classifier and therefore $C_{ID} = 0$. It can also be seen that the composition function $Y = f(C)$ always results in the same C_{ID} to its inverse $Y = \neg f(C)$, for example f_3 and f_{14} or f_4 and f_{13}. Assuming a fixed cost analysis, we can compute the likelihood ratio $l(C)$ (Section 2.1, Eq. (1)) for each state of C and consider them as possible values of τ to derive the composition function that minimizes the expected cost (Table 2 a, index of τ_x sorted by descending $l(C)$) (Gu et al., 2008). Since the base-rate is constant, each τ directly relates to a specific fixed cost ratio. More importantly, it is known that all systems, which achieve an identical expected cost, can be found on a straight line in the ROC plot which has the slope τ (Cardenas et al., 2006). For example, if we consider $\tau_3 = 1.11$, the LRT (Eq. (1)) results in the two functions of identical minimal cost, f_4 and f_{12} (Gu et al., 2008). All points which extend the red line that is marked as τ_3 in Fig. 7 a, result in the identical expected cost at the respective cost ratio; all points below correspond to a higher expected cost (Cardenas et al., 2006). Evaluating the values of the remaining τ_x or one value of each in-



(a) $p = 0.20$ (b) $p = 0.001$ (c) $p = 0.0001$

Fig. 7. ROC curve analysis with isolines for C_{ID} at different base rates.

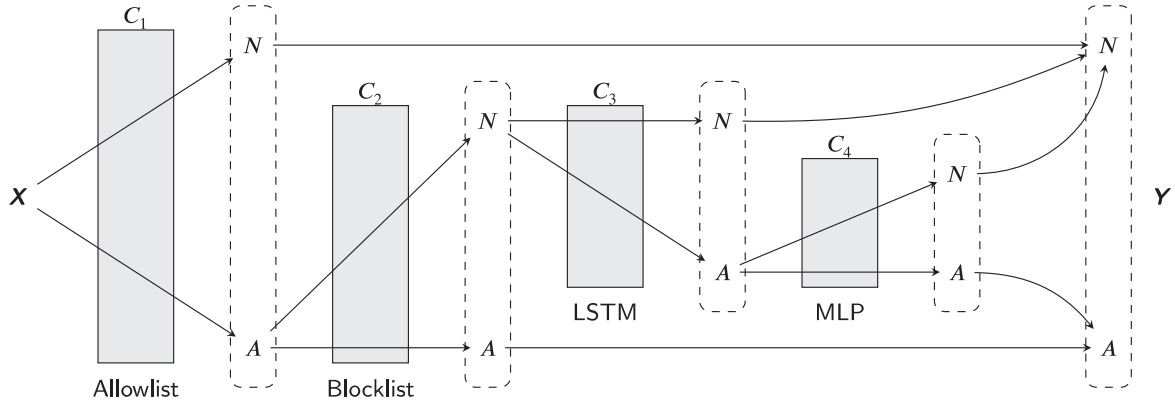


Fig. 8. Evaluation example system (sequential representation): The allowlist shall increase the effective base-rate at the main NIDS (LSTM) and the false alarm filter (MLP) shall increase the PPV at low base-rates.

terval (τ_x, τ_{x+1}), leads to the ideal composition functions for any fixed cost f_3, f_4 and f_{12}, which are shown on the red line that covers all remaining points (Gu et al., 2008).

Since we know that the isolines for a fixed cost are straight in the ROC plot while the isolines for C_{ID} are always *concave* if TPR > FPR for any base-rate, we know that the composition function which maximizes C_{ID} also minimizes some fixed cost ratio. This is equivalent to the previous observation of Cardenas et al. (2006). Visually, it means that for TPR > FPR the area below the straight line between any two points in a ROC plot corresponds to an area of lower C_{ID} than at least of the two points. Therefore, the ideal composition form an information-theoretic perspective must be one of the functions that results from the LRT approach of Gu et al. (2008). In this case, the ideal composition function is f_{12} with $C_{ID} = 0.196$ at the base-rate $p = 0.2$.

Since a ROC plot is in itself base-rate independent, the composition functions of Table 2 maintain their position by considering the results for lower base-rates as shown in Fig. 7b and 7 c. However, C_{ID} is base-rate dependent which causes the isolines to change accordingly (notice that the FPR axis has been limited for visualizing the isolines of C_{ID}). It can be seen as expected that the same point in a ROC plot corresponds to a lower C_{ID} at lower base-rates. This changes the ideal composition function at the base-rates $p = 0.001$ and $p = 0.0001$ to f_4 with $C_{ID} = 0.081$ and $C_{ID} = 0.063$, respectively.

This highlights another time the close connection between C_{ID} and the expected cost, even though both metrics have very different properties. It also confirms that the composition function which minimized the composition loss L_Y can be found using the LRT with one value of each interval (τ_x, τ_{x+1}).

4. Evaluation

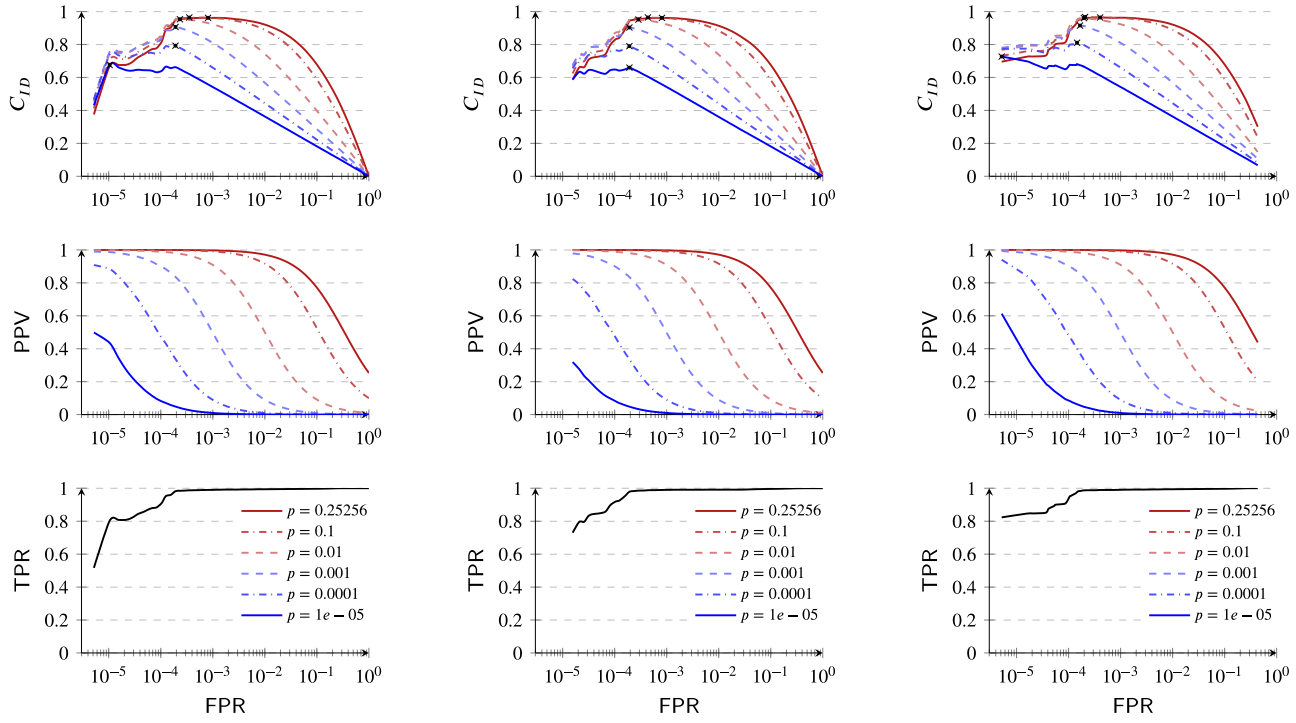
This section aims to demonstrate the application of the Information-Theoretic Framework. It will be used to design, fine-tune and evaluate an exemplary composed NIDS using systems described in the literature and an open data set. The system performance will be analyzed for a variety of base-rates and the operation points will be optimized for the composition function. Additionally, the system performance will be attributed to its individual components to better understand the impact of an evasion attempt with adversarial examples. Please notice that the used NIDSs, data set, threat model and attack method only serve as simple examples for the demonstration of the framework.

The considered system is composed of four classifiers as shown in Fig. 8. It consists of a blocklist (C_2) and allowlist (C_1), which shall increase the effective base-rate at the primary detection method, due to the high class imbalance in practical environments. The primary detection method is a Long Short-Term Memory (LSTM) model (C_3) that was trained by Hartl et al. (2020). Additionally, a Multilayer Perceptron (MLP) will be used as False-Alarm Filter (C_4) to increase the final PPV at low base-rates. The resulting sequential composition, as shown in Fig. 8, can be expressed by the composition function $Y = C_1 \wedge (C_2 \vee (C_3 \wedge C_4))$ in the framework of Section 3.

For the analysis, we used the *CIC-IDS-2017* data set from the Canadian Institute of Cybersecurity (Sharafaldin et al., 2018) and the pre-processing by Hartl et al. (2020) with an identical split of the training and test set to match the setup of the used LSTM model. This provides the feature representation R_1, which presents each flow as sequence of packets and is used by the classifier C_3. The additional classifiers C_1, C_2 and C_4 use an aggregated feature representation R_2 of each flow, which is similar to the pre-processing of Bachl et al. (2019). We use a second feature representation because of both, practical and educational aspects. A variety of feature representations is expected to increase the chance of observing independent classification redundancies during an evasion attempt and an aggregated feature space specifically could provide more robustness against the considered perturbation operations in Section 4.2. An overview of both feature representations can be found in Appendix 4 and the used unit-of-analysis in all following evaluations is a flow as defined by the used data set. Since the used test set has an original base-rate of about 25%, the results will be projected to lower base-rates under the assumption of a constant TPR and FPR to study the impact of the base-rate fallacy in practical environments. This assumption by projecting the results to lower base-rates will be discussed further in Section 5.

To generate a simple rule set that shall serve as blocklist and allowlist, a first validation set has been split off the training data (1/3). A decision tree was generated using Scikit-learn (Pedregosa et al., 2011) on the feature representation R_2. The validation set was then used to isolate individual decision rules with at least 100 matches and an error-rate below 10^{-4}. This resulted in a naive allowlist of 16 rules (C_1) and a naive blocklist of 10 rules (C_2).

The MLP (C_4) was generated using Pytorch (Paszke et al., 2017). It has 5 fully connected layers of each 512 neurons with Rectified Linear Unit (ReLU) activation function and 0.2 dropout probability. It uses the feature representation R_2 and was trained with binary cross entropy as loss function using the optimizer Adam (Kingma and Ba, 2015). This resembles a model by Bachl et al. (2019) on the same data set.



(a) LSTM results (C_3) (b) MLP results (C_4) (c) Composed results ($C_3 \wedge C_4$)

Fig. 9. System evaluation viewing TPR, PPV and C_{ID} depending on the allowed FPR. The $\wedge$-composition only achieves a noticeable performance difference for low base-rates and FPR ranges.

4.1. System performance analysis

A second validation set has been split off from the test set (1/3) for an operation point analysis and optimization. This subset is excluded from the later results on the test set.

Performance overview: The rule-based methods have a single operation point, which results in (FPR= 91.68%, TPR= 99.995%) for the allowlist (C_1) and (FPR= 0.001%, TPR= 25.40%) for the blocklist (C_2) on the validation set. This corresponds to a $C_{ID_1} = 0.007$ and $C_{ID_2} = 0.185$ at an example base-rate of 10^{-5}, respectively. It is expected that the allowlist has a lower C_{ID} than the blocklist as it only serves to correctly identify a subset of benign samples which have a lower entropy compared to attack samples.

The performance evaluation of the classifiers C_3 and C_4 on the validation set can be seen for a variety of operation points and base-rates in Fig. 9. It shows the TPR, PPV and C_{ID} depending on the selected FPR on a logarithmic scale. It can be seen that both classifiers individually achieve comparable results (Fig. 9a and 9 b). It can also be seen as expected that the maximal C_{ID} and corresponding PPV decrease significantly with reducing base-rate. The $\wedge$-composition (Fig. 9c) shows an advantage over its components with decreasing base-rates. It outperforms both individual classifiers with a higher TPR, PPV and C_{ID} at FPRs below 10^{-4}.

Operation point analysis: To further visualize the impact of composing $C_3 \wedge C_4$ the resulting C_{ID} on the validation set is shown for a variety of detection threshold combinations at a base-rate of 10^{-5} in Fig. 10. The maximal C_{ID} for the individual and composed system have been marked to highlight how the ideal operation point changes as a result of the composition and how the composed C_{ID} of 0.73 exceeds the maximal 0.70 and 0.67 of its individual systems. The ideal detection thresholds would change further by adding the classifiers C_1 or C_2 and the operation point analysis could be done independent of the specific composition by viewing $C_{C_{(3,4)}} - L_{Y_{\min}}$. However, for simplicity of this example, the detection thresholds that maximized the operation point of $C_3 \wedge C_4$ will be used for all following analyses at the same base-rate.

Feature representation analysis: With the adjusted operation points, the test set will be analyzed. To include the feature representation in the performance decomposition, the corresponding feature representation capabilities and possible gain will be analyzed. Since the feature representation $R_2 = f(R_1(x))$ can be expressed as function of R_1, it follows from the data processing lemma that its information gain must be zero ($G_{R(1,2)} = 0$). Both representations contain at least one probabilistic feature, which is

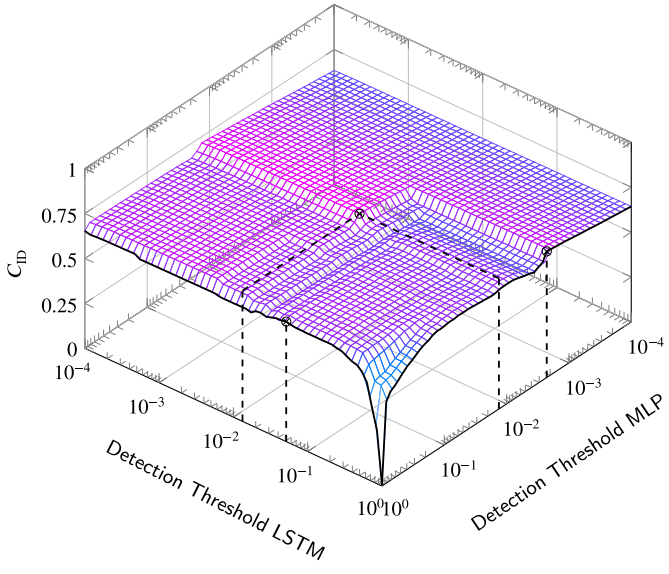


Fig. 10. Operation point analysis of $C_3 \wedge C_4$ at $p = 1e-05$.

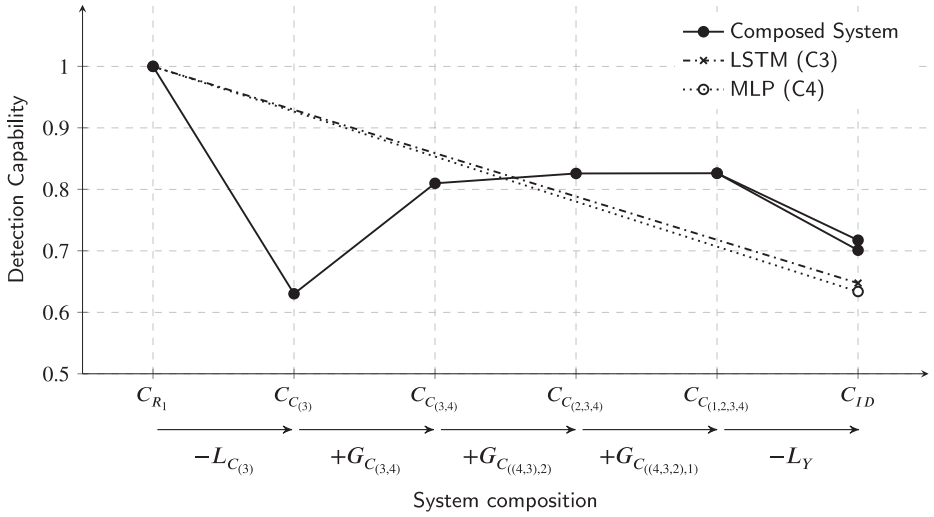


Fig. 11. System performance decomposition at $p = 10^{-5}$.

the inter-arrival time. While it would be possible to estimate a distribution for the timing jitter, it is easier to instead specify bounds for their feature representation capabilities. The upper bound was computed by specifying two samples as equal if they match exactly (including the inter-arrival time) and the lower bound by ignoring any timing deviations between samples. This gave identical results for both feature representations, leading to $0.9998 \leq C_{R_2} \leq C_{R_1} \leq 0.9999$. Most undistinguishable samples to the normal traffic are attack classes labeled Denial of Service (DoS), Infiltration and Web Attack, which could indicate that the representation may benefit from additional traffic-features with statistics related to past connections or content features if they were available.

System performance decomposition: The system performance decomposition for the test set results can be seen in Fig. 11, which shows the different capabilities of the system based on the classifier C_3. The respective gains and losses can be seen as transitions. The classifier C_3 causes an initial classification information loss of $L_{C_{(3)}} = 0.370$ from the feature representation capability C_{R_1}. This is more than if it would operate at its individually best operation point, but leads to a better overall system performance as indicated by Fig. 10. Adding the classifier C_4 creates a classification gain of $G_{C(3,4)} = 0.180$. Since this is below its classification capability with $C_{C_3} = 0.630$, it could provide redundancy for C_3. Adding the blocklist and allowlist only causes minimal increases in the classification capability with gains of 0.016 and 0.0004, providing additional redundancy. The composition function $Y = C_1 \wedge (C_2 \vee (C_3 \wedge C_4))$ of Fig. 8 causes a composition loss of $L_Y = 0.125$, leading to a final performance of $C_{ID} = 0.701$. This exceeds the performance of all individual components. Based on the joint probabilities of $C_{C_{(1,2,3,4)}}$, it can be found that the alternative composition function $Y = (C_1 \wedge C_2 \wedge C_4) \vee (\neg C_2 \wedge C_3 \wedge C_4)$ would lead retrospectively to the maximal system performance of $C_{ID} = 0.717$ as upper bound on the achievable performance for these classifiers and thresholds.

4.2. System robustness analysis

After the performance has been attributed to its individual components, the robustnuss of the classifiers C_3 and C_4 is analyzed on the test set together with its impact on the overall system. It will be considered a minimalistic threat model, which only serves as example to demonstrate parts of the analysis out of Section 3.3. The used threat model and attack method are based on the ℓ_1-distance attack by Hartl et al. (2020), with modification for multiple classifiers.

The threat model considers that the adversary has knowledge of the detection models (white-box scenario) and the allowed perturbation operations only consist of delaying packets and increasing their size. These perturbations can only be applied to packets in the forward direction within logical constraints, such as a maximal inter-arrival time or maximal packet size. This aims on deriving misclassified flows without affecting their functionality. As simple attack budget will be considered that the adversary could increase its bandwidth and attack time by up to 25% compared to its capabilities in the original test set. Additionally will be assumed that the NIDS operates at a constant operation point, since a dynamic adjustment based on the alarm rate could be exploited by generating false alarms.

Ideal robustness limitations: Based on the perturbation operations of the threat model, a reachabiliy analysis can be performed to estimate the maximal performance of an ideal robust classifier on the test set. However, to simplify the analysis, a lower bound is computed instead by defining two samples as reachable if they are identical while ignoring their timing and packet sizes. The removal of the timing and size is the simplest pre-processing which maps all derived samples under the naive threat model to the same feature vector, ensuring the lower bound. This analysis leads to the new bounds of $0.991 \leq C'_{R_2} \leq 0.9999$ and $0.992 \leq C'_{R_1} \leq 0.9999$, where the undistungishable samples are caused by the same attack classes as before. This highlights that a robust oracle would exist and indicates that the evasion can not be independent of the used classifiers.

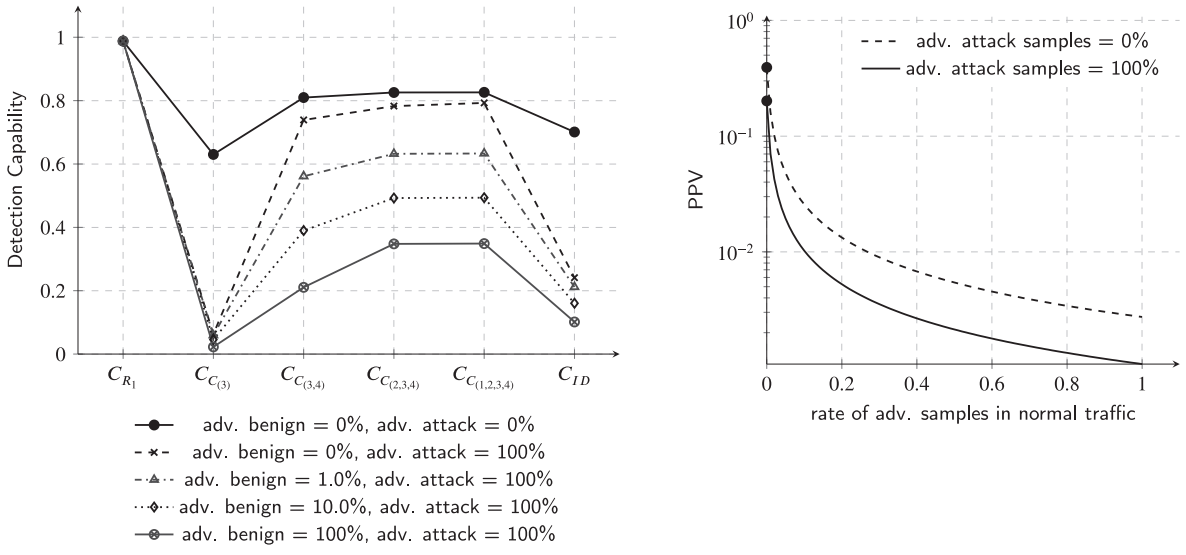
Adversarial attack method: The considered attack for evaluating the classifier robustness is based on the Carlini-Wagner method (Carlini and Wagner, 2017). The adversarial examples are generated based on the optimization objectives of Table 3, which solution is approximated using the Adam optimizer (Kingma and Ba, 2015). It aims to minimize the difference between the logit output of the LSTM (T_3) and MLP (T_4) to the desired logit output υ_i while penalizing increases in time (Δt) and size (Δs) between the original sample $\vec{x}$ and adversarial sample $\vec{x^*}$ with the trade-off parameters ε. As desired output for attack samples has been selected υ_n just below the operation points and υ'_n for normal samples just above the selected operation points.

The ε-parameters have been adjusted such that the average inter-arrival time was increased by 23.0% and the average package size by 8.5% to ensure staying within the specified attack budget.

By using the described method, adversarial examples have been generated for each flow in the test set. Its impact on the individ-

Table 3
Used optimization objectives for generating adversarial examples.

Targeted error type	Optimization objective
False Negatives	$\max(T_3(R_1(\vec{x}^*)) - \upsilon_3, 0) + \max(T_4(R_2(\vec{x}^*)) - \upsilon_4, 0) + \varepsilon_t \cdot \Delta t(\vec{x}, \vec{x}^*) + \varepsilon_s \cdot \Delta s(\vec{x}, \vec{x}^*)$
False Positives	$\max(\upsilon'_3 - T_3(R_1(\vec{x}^*)), 0) + \max(\upsilon'_4 - T_4(R_2(\vec{x}^*)), 0) + \varepsilon_t \cdot \Delta t(\vec{x}, \vec{x}^*) + \varepsilon_s \cdot \Delta s(\vec{x}, \vec{x}^*)$



(a) Performance decomposition

(b) Impact of adversarial false-positives

Fig. 12. Robustness analysis at $p = 10^{-5}$ of $Y = C_1 \wedge (C_2 \vee (C_3 \wedge C_4))$.

ual classifiers can be seen in Table 4, which compared the performance on the original to the adversarial test set at a base-rate of $p = 10^{-5}$. The used attack method has a high success rate by generating false-negatives against the used LSTM, but increased the detection rate of the used MLP. The MLP was on the other hand more susceptible to the generation of false-positives than the LSTM. Table 4 also highlights that the performance of the naive block- and allowlist decreased, even though they have not been considered during the generation of the adversarial examples. This results in a low PPV and C_{ID} for all classifiers.

Evasion attempt - system performance decomposition: Fig. 12a attributes the overall system performance to the classifiers in Table 4. Fig. 12 a compares the system performance for different assumptions of which part of the traffic are under adversarial control and could be perturbed. For attack samples has been considered that either all of them have an adversarial perturbation (adv. attack = 100%) or none of them (adv. attack = 0%). For the benign samples has been considered that 0%, 1%, 10% or 100% of the normal traffic are under control of the adversary (adv. benign) and could thus be adversarially perturbed to cause false alarms.

Table 4
Adversarial attack performance at $p = 10^{-5}$.

model	test set	TPR	FPR	PPV	C_{ID}
Allowlist (C1)	normal	99.998%	91.68%	0.001%	0.007
Allowlist (C1)	adversarial	94.54%	90.02%	0.001%	0.001
Blocklist (C2)	normal	25.30%	0.001%	19.55%	0.184
Blocklist (C2)	adversarial	19.72%	0.057%	0.343%	0.078
LSTM (C3)	original	95.94%	0.022%	4.16%	0.630
LSTM (C3)	adversarial	14.21%	0.808%	0.018%	0.023
MLP (C4)	original	84.39%	0.007%	11.48%	0.611
MLP (C4)	adversarial	97.58%	6.58%	0.015%	0.203

While the performance of C_1 decreases as a result of the adversarial attack samples, it can be seen an increase in the information gain of C_2. This highlights that the redundancy of C_2 could be utilized at the right composition function. For example, in the case of adversarially perturbed attacks and normal benign traffic (adv. benign = 0.0%, adv. attack = 100% in Fig. 12 a), the overall system performance decreased due to the composition loss to $C'_{ID} = 0.24$ while an alternative composition would exist that maintains $C'_{ID} = 0.74$. As also the information gain of C_2 decreases with increasing adversarial perturbations among benign flows, it can be seen that the information gain of the blocklist C_2 slightly increases. Interestingly, this ends in a synergistic affect where $C'_{C_2} = 0.078 < C'_{C_{((4,3),2)}} = 0.137$ at the full adversarial test set (100% adv. benign, 100% adv. attack in Fig. 12 a). These redundancies and synergies are however not fully utilized in the detection outcome due to the static composition function. The retrospectively ideal composition function for the full adversarial test set would result in $C'_{ID} = 0.206$, which is only slightly above all individual classifiers but higher than the resulting $C'_{ID} = 0.101$ of Fig. 12 a. That the used LSTM is more susceptible to adversarial perturbations than the MLP could be expected from the used feature representation, as the aggregated feature space (R_2) provides a smaller attack surface for the operations of the threat model compared to R_1.

Fig. 12 b further highlights the impact of adversarial perturbations from normal traffic on the PPV of the composed system due to false alarms. It considers that either all or non of the attack samples have adversarial perturbations (adv. attack samples 100% and 0%) and shows the PPV depending on which fraction of the normal traffic can be adversarially perturbed. It can be seen that a small percentage of adversarially perturbed benign flows have a high impact on the PPV. These results could be amplified further by assuming that the adversary would preferably generate selected adversarial benign samples that have a high likelihood of causing false alarms at one, multiple or the composed system.

Overall, it can be seen that redundancies between the classifiers could be utilized and possibly improve the robustness of all individual classifiers if the composition function was dynamically adjusted based on feedback to the false alarms. However, the results also highlight that both, the specific individual and composed system, are vulnerable to the threat model which may render them practically unusable.

5. Discussion

The ITF provides an objective skill metric with additive properties to attribute the overall system performance to its individual components. The evaluation demonstrated that maximizing the performance of one component individually does not necessarily maximize the overall system performance. It could also seen that a vulnerable component might require the dynamic adjustment of the composition function to maintain the system performance. Finally, it has been possible to identify classification redundancies and evaluate their dependence by using an example evasion attempt. This can provide new insights to the robustness of composed detection systems and by studying the impact of adversarial attacks on complex systems.

Composed systems can be analyzed from a cost and information-theoretic perspective. Both approaches are related, data-driven, provide an evaluation metric for system comparisons and enable the fine-tuning of a system by optimizing its operation points or composition function. However, both metrics provide very different properties which gives them complementing use cases.

The expected cost directly provides a practical meaning to the evaluation results and enables incorporating operational costs into the analysis. However, the results can not be attributed to the individual components for identifying system limitations and the evaluation metric might rank for example a random classifier over a system with skill. Moreover, the cost ratio might be estimated subjectively and bias the analysis results. While some of these aspects might be undesirable in an objective analysis, they also directly relate to the practical *value* which the system provides to a specific user. This makes the expected cost a suitable tool to select and optimize a NIDS for a specific and known deployment environment.

The presented information-theoretic approach provides an objective analysis metric, which attempts to measure the *skill* of a system rather than the value and can avoid vanishing results at low base-rates. The approach enables an attribution of the overall system performance to its individual components, which provides important insights on the system limitations for further improvements. Moreover, the information-theoretic perspective gives the opportunity to find indications of classification synergies and redundancies, as seen during the evaluation of this work. This could be especially useful for deriving robustness requirements on the different classifiers, since it can indicate which parts of the performance rely on which components and might be volatile in case of their evasion. Since the framework is based on classical information theory, it does not provide a direct measure of redundancies about a target variable. However, changes in the system redundancy can be observed indirectly by the difference between its capability and gain (normalized co-information). This highlights that applying a partial information decomposition in future work on this framework could provide further insights to understand the robustness of a composed NIDS and guide the design of resilient detection systems. These properties of the composed ITF make it a suitable tool for an objective and fine-grained analysis that is independent of a specific deployment, or for studying the dynamics and robustness of composed NIDSs.

The analysis of the ITF is based on a data driven approach which leads to a number of limitations since the evaluation inherits all shortcomings of the used data set. For example, the results are not expected to generalize to other networks or changes over time. Similarly, the attacks and background traffic of the used data set may not be representative for the targeted application or its environment. The direct analysis as done in Section 4 also does not indicate any anomaly or novelty detection capability, since this would require leaving out attack classes during training or performing additional benchmarks from the anomaly detection area. This work focused on a binary detection problem, but the random variables of the framework could be extended for incorporating different attack classes. It will be part of future work to analyze the framework results for multiple data sets and models.

Since the original base-rate of the used data set was at about 25%, the results have been projected to lower base-rates to consider the impact of the base-rate fallacy in the system evaluation. One key assumption for studying this effect was that the TPR and FPR are base-rate independent. While this holds at first sight, it also implies that the distribution of attack classes would stay identical and that the attacks would have no side-effects on the features of other flows which are impacted by the base-rate scaling. To which extent both of these assumptions hold depends on the specific application, data set, and used feature representation. However, scaling the base-rate can provide meaningful estimations for the expected system performance that are closer to their deployment scenario.

Finally, the performed robustness analysis only provides a lower bound on the achievable performance degradation, since stronger threat models, attack methods or parameters might exist. For performing meaningful robustness evaluations, it would be required to further study different attack classes and which perturbation operations can be performed on the different layers of the network stack without affecting a flow's functionality. Nevertheless, using adversarial examples provides an objective analysis approach which does not require subjective estimations that could bias the final results based on prior beliefs.

6. Conclusion

Composed detection systems might gain increasing importance in the future, since they can address the rising challenge from adversarial machine learning with the system diversification. The Information-Theoretic Framework of Gu et al. (2006b) provided the key advantage of an objective evaluation metric with additive properties, but has not been suited for the analysis of such composed NIDSs. The presented framework resolves this limitation and provides deeper insights on how to improve the performance and robustness of a specific design.

The information-theoretic analysis of composed detection systems enables identifying the importance of each component on the overall performance, which can be used to derive requirements for the robustness and dependencies between classifiers for a resilient system design. Similarly, it can be used for investigating how to design a resilient NIDS based on the available detection methods and known evasion approaches or to understand the detailed impact of adversarial attacks on complex systems. This makes the presented framework a valuable tool for the design, analysis and comparison of modern NIDSs.

Funding

This work was supported by the Swedish Civil Contingencies Agency (MSB) through the project RIOT grant number MSB 2018-12526.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Undistinguishability definition example

The following example shall highlight how the data processing lemma of the ITF could be violated and motivate the intuitive modification of the definition for undistinguishability. Consider twenty samples of which eleven are attacks (A) and nine are benign (N). Assume a feature representation under which three benign samples have an identical feature vector to one attack sample and assume that five other attack samples have an identical feature vector to one other benign sample. This leads to the joint probabilities shown in the Tables A.1 at the original base-rate.

Table A.1 a provides a possible oracle classifier for the feature representation which leads to $C_{ID} \approx 0.53$. Table A.1 b uses the definition for undistinguishability of Eq. (2), leading to $C_R \approx 0.51$. This causes the unexpected contradiction that $C_R \not\geq C_{ID}$, which would violate the data-processing lemma of the ITF. Table A.1 b uses the definition of Eq. (6), leading to $C_R \approx 0.64$ and maintaining $C_R \geq C_{ID}$.

Also notice that changing the definition of undistinguishability changes the classification loss - in this case from $L_C \approx -0.02 \not\geq 0$ to $L_C \approx 0.11 \geq 0$.

Appendix B. Additivity of the evaluation metrics

The definitions the composed ITF lead to the additive properties that $C_{ID} = C_R - L_C - L_Y = C_C - L_Y$, $C_{R_{(1,2)}} = C_{R_1} + G_{R(1,2)}$ and $C_{C_{(1,2)}} = C_{C_1} + G_{C(1,2)}$, which will be shown below:

Theorem 1. $C_{ID} = C_R - L_C - L_Y = C_C - L_Y$.

Proof. Similarly to the proof by Gu et al. (2006b), the properties of a Markov chain can be utilized. The random variable Y is conditionally independent from X given Z or C, such that $I(X;Y|C,Z) = I(X;Y|Z) = I(X;Y|C) = 0$. Similarly the random variable C is conditionally independent from X given Z, leading to $I(X;C|Z) = 0$.

Mutual information can be expanded in different ways by using the chain rule.

$$\begin{aligned} I(X;Z,C,Y) &= I(X;C) + I(X;Z|C) + I(X;Y|C,Z) \\ &= I(X;C) + I(X;Y|C) + I(X;Z|C,Y) \end{aligned} \tag{B.1}$$

By setting both versions equal and knowing the conditional independence, it directly follows that $I(X;Z|C,Y) = I(X;Z|C)$

Moreover can mutual information be expanded to:

$$\begin{aligned} I(X;Z,C,Y) &= I(X;Z) + I(X;C|Z) + I(X;Y|C,Z) \\ &= I(X;Y) + I(X;C|Y) + I(X;Z|C,Y) \end{aligned} \tag{B.2}$$

Again by setting both versions equal, knowing the conditional independence and that $I(X;Z|C,Y) = I(X;Z|C)$, it directly follows that $I(X;Y) = C_{ID} \cdot H(X) = I(X;Z) - I(X;Z|C) - I(X;C|Y) = (C_R - L_C - L_Y) \cdot H(X)$.

By expanding mutual information one last time to,...

$$\begin{aligned} I(X;Z,C) &= I(X;Z) + I(X;C|Z) \\ &= I(X;C) + I(X;Z|C) \end{aligned} \tag{B.3}$$

... setting both versions equal and knowing the conditional independence, it follows that $I(X;C) = C_C \cdot H(X) = I(X;Z) - I(X;Z|C) = (C_R - L_C) \cdot H(X)$ □

Theorem 2. *$C_{C_{(1,2)}} = C_{C_1} + G_{C(1,2)}$ and $C_{R_{(1,2)}} = C_{R_1} + G_{R(1,2)}$ where C_1 and C_2 are two classifiers $C = (C_1, C_2)$ and Z_1 and Z_2 are two feature representations $Z = (Z_1, Z_2)$*

Proof. Both cases directly follow from expanding the mutual information:

$$I(X;C_1..C_n) = I(X;C_1..C_{n-1}) + I(X;C_n|C_1..C_{n-1}) \tag{B.4}$$

This demonstrates that $I(X;C_1..C_n) = C_{C_{(1..n)}} \cdot H(X) = I(X;C_1..C_{n-1}) + I(X;C_n|C_1..C_{n-1}) = (C_{C_{(1..n-1)}} + G_{(1..n-1,n)}) \cdot H(X)$.
The example above uses only two classifiers such that $n = 2$ and the approach can equally be applied to the definitions of the feature representation capability and gain. □

Appendix C. Composition function example

The following example shall highlight that C_{ID} is not necessarily maximized by minimizing P_e. Consider two classifiers which result in the joint probabilities shown in Table C.1 a at a base-rate $p = 0.0005$. Using Bayes decision rule to define the composition function Y_1 would result in a minimal P_e but only $C_{ID} \approx 0.15$ as shown in Table C.1 b. However, an alternative composition Y_2 with worse P_e can achieve a higher $C_{ID} \approx 0.75$ as shown in Table C.1 c.

Another example can be generated by modifying Table C.1 a such that $P(X = N|C = (A,A)) > P(X = A|C = (A,A))$. In this case, the composition of lowest P_e would be equivalent to a constant classifier and receive the worst score of $C_{ID} = 0$.

Table A.1
Undistinguishability example.

Y \ X	A	N
N	0.05	0.4
A	0.5	0.05

(a) Possible classification output

Z \ X	A	N
N	0	0.25
U	0.3	0.2
A	0.25	0

(b) Undistinguishability based on Equation 2

Z \ X	A	N
N	0	0.25
U_1	0.05	0.15
U_2	0.25	0.05
A	0.25	0

(c) Undistinguishability based on Equation 6

Table C.1
Composition function examples.

C \ X	N	A	$Y_1 = f_1(C)$	$Y_2 = f_2(C)$
(N, N)	0.0005	0.0002	N	A
(A, N)	0.0005	0.0002	N	A
(N, A)	0.99849	1e-5	N	N
(A, A)	1e-5	9e-5	A	A

(a) Joint probabilities $C_C \approx 0.77$

Y_1 \ X	N	A
N	0.99949	0.00041
A	1e-5	9e-5

(b) Bayes composition $C_{ID} \approx 0.15$, $P_e = 0.00042$

Y_2 \ X	N	A
N	0.99849	1e-5
A	0.00101	0.00049

(c) Alternative composition $C_{ID} \approx 0.75$, $P_e = 0.00102$

Table D.1
Used feature representations in the system example.

	R_1	R_2
representation level	packet sequence	aggregated flow
number of features	15 per packet	40
Source port	✓	✓
Destination port	✓	✓
Protocol ID	✓	✓
Packet length	✓	Min/Max/Mean/Std - forward/backward
Inter-arrival time	✓	Min/Max/Mean/Std - forward/backward
Direction	✓	count - forward/backward/total
Flags*	✓	count - forward/backward

Flags* = (SYN, FIN, RST, PSH, ACK, URG, ECE, CWR, NS)

Appendix D. Feature representations of the evaluation

Table D.1 gives an overview the two feature representations that are used by the classifiers in the evaluation. The representation R_1 directly follows from the pre-processed data set by Hartl et al. (2020) and the representation R_2 resembles the representation used by Bachl et al. (2019).

Appendix E. List of symbols

Table E.1 provides an overview of the used symbols with description.

Table E.1
List of symbols.

Symbol	Description
Fundamental evaluation metrics:	
p	base-rate, $P(I)$
α, FPR	false positive/false alarm rate, $P(A\|\neg I)$
β, FNR	false negative/missed alarm rate, $P(\neg A\|I)$
TPR	true positive/detection rate, $P(A\|I)$
TNR	true negative rate, $P(\neg A\|\neg I)$
PPV	positive predictive value, $P(I\|A)$
NPV	negative predictive value, $P(\neg I\|\neg A)$
FDR	false discovery rate, $P(\neg I\|A)$
FOR	false omission rate, $P(I\|\neg A)$
P_e	error probability, $P(X \neq Y)$
b	ratio of false discovery rate to false omission rate, FDR/FOR
Cost analysis and likelihood ratio test:	
c_α	cost of a false positive/false alarm
c_β	cost of a false negative/missed intrusion
c_{op}	cost objective of Gaffney and Ulvila (2001)
c_{exp}	expected cost
c	cost ratio c_β / c_α
$l(\vec{A})$	likelihood ratio, $P(\vec{A}\|I)/P(\vec{A}\|\neg I)$
τ	decision threshold for the likelihood ratio, $c_\alpha(1-p)/(c_\beta p)$

(*continued on next page*)

Table E.1 (*continued*)

Symbol	Description
Information-Theoretic Framework:	
{N,U,A}	possible states: normal/benign (N), undistinguishable (U), anomalous/attack (A)
D_i	input data stream i
F_i	feature representation of data stream i, $R(D_i)$
O_{NIDS}	oracle NIDS
X_{FA}	state of the sample (label) at the false alarm filter input
Y_{FA}	state of the false alarm filter output
RC_{FA}	false alarm reduction capability, $I(X_{\text{FA}};Y_{\text{FA}})/H(X_{\text{FA}})$
X	state of the sample (label), $O_{\text{NIDS}}(D_i)$
Z	state of the feature representation, $L_R(F_i)$
C	state of the classification outputs
Y	state of the system output
C_{ID}	intrusion detection capability, $I(X;Y)/H(X)$; under adversarial attack C'_{ID}
C_C	classification capability, $I(X;C)/H(X)$; under adversarial attack C'_C
C_R	feature representation capability, $I(X;Z)/H(X)$; under adversarial attack C'_R
L_Y	composition information loss, $I(X;C\|Y)/H(X)$; under adversarial attack L'_Y
L_C	classification information loss, $I(X;Z\|C)/H(X)$; under adversarial attack L'_C
$G_{C(1..n,n+1)}$	classification information gain, $I(X;C_{n+1}\|C_{1..n})/H(X)$; under adversarial attack $G'_{C(1..n,n+1)}$
$G_{R(1..n,n+1)}$	representation information gain, $I(X;Z_{n+1}\|Z_{1..n})/H(X)$; under adversarial attack $G'_{R(1..n,n+1)}$
Adversarial robustness:	
$(\delta_p,\delta_\alpha,\delta_\beta)$-intruder	base-rate interval (δ_p), false alarm success rate (δ_α) and evasion probability (δ_β)
$\vec{x}$	original data sample
$\vec{x^*}$	adversarial perturbed version of sample x
$\delta_{\vec{x}}$	adversarial sample perturbation, $\vec{x^*}-\vec{x}$
ℓ_p	distance measure between samples
υ_n	targeted (logit) output for false negatives at classifier n
υ'_n	targeted (logit) output for false positives at classifier n
ε_s	weight of the size increase in the optimization objective
ε_t	weight of the time increase in the optimization objective
T_n	logit output of classifier n
$R_i(x)$	feature representation i of sample x

CRediT authorship contribution statement

Tobias Mages: Conceptualization, Methodology, Software, Writing – original draft. **Magnus Almgren:** Writing – review & editing, Supervision. **Christian Rohner:** Conceptualization, Methodology, Writing – review & editing, Supervision.

References

Axelsson, S., 1999. The base-rate fallacy and its implications for the difficulty of intrusion detection. In: Proceedings of the 6th ACM Conference on Computer and Communications Security. Association for Computing Machinery, New York, NY, USA, p. 17. doi:10.1145/319709.319710.

Bachl, M., Hartl, A., Fabini, J., Zseby, T., 2019. Walling up backdoors in intrusion detection systems. In: Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks. Association for Computing Machinery, New York, NY, USA, p. 813. doi:10.1145/3359992.3366638.

Birnbach, S., Eberz, S., Martinovic, I., 2019. Peeves: physical event verification in smart homes. In: Proceedings of the ACM SIGSAC Conference on Computer and Communications Security. Association for Computing Machinery, New York, NY, USA, p. 14551467. doi:10.1145/3319535.3354254.

Bossomaier, T., Barnett, L., Harré, M., Lizier, J.T., 2016. An Introduction to Transfer Entropy. Cham: Springer International Publishing doi:10.1007/978-3-319-43222-9.

Cardenas, A.A., Baras, J.S., Seamon, K., 2006. A framework for the evaluation of intrusion detection systems. In: Proceedings of the IEEE Symposium on Security and Privacy (S P'06), pp. 15pp.–77. doi:10.1109/SP.2006.2.

Carlini, N., Athalye, A., Papernot, N., Brendel, W., Rauber, J., Tsipras, D., Goodfellow, I., Madry, A., Kurakin, A., 2019. On evaluating adversarial robustness. 1902.06705.

Carlini, N., Wagner, D., 2017. Towards evaluating the robustness of neural networks. In: Proceedings of the IEEE Symposium on Security and Privacy (SP), pp. 39–57. doi:10.1109/SP.2017.49.

Durst, R., Champion, T., Witten, B., Miller, E., Spagnuolo, L., 1999. Testing and evaluating computer intrusion detection systems. Commun. ACM 42 (7), 5361. doi:10.1145/306549.306571.

Fano, R.M., 1961. Transmission of Information: A Statistical Theory of Communications. The MIT Press.

Gaffney, J.E., Ulvila, J.W., 2001. Evaluation of intrusion detectors: a decision theory approach. In: Proceedings of the IEEE Symposium on Security and Privacy S P 2001, pp. 50–61. doi:10.1109/SECPRI.2001.924287.

Gandin, L.S., Murphy, A.H., 1992. Equitable skill scores for categorical forecasts. Mon. Weather Rev. 120 (2), 361–370. doi:10.1175/1520-0493(1992)120<0361:ESSFCF>2.0.CO;2.

Giacinto, G., Roli, F., 2002. Pattern Recognition for Intrusion Detection in Computer Networks. Springer US, Boston, MA, pp. 195–218. doi:10.1007/978-1-4613-0231-5_8. chapter 3

Goodfellow, I. J., Shlens, J., Szegedy, C., 2015. Explaining and harnessing adversarial examples. 1412.6572.

Gu, G., Cárdenas, A.A., Lee, W., 2008. Principled reasoning and practical applications of alert fusion in intrusion detection systems. In: Proceedings of the ACM Symposium on Information, Computer and Communications Security. Association for Computing Machinery, New York, NY, USA, p. 136147. doi:10.1145/1368310.1368332.

Gu, G., Fogla, P., Dagon, D., Lee, W., Skorić, B., 2006a. Measuring intrusion detection capability: an information-theoretic approach. In: Proceedings of the ACM Symposium on Information, Computer and Communications Security. Association for Computing Machinery, New York, NY, USA, p. 90101. doi:10.1145/1128817.1128834.

Gu, G., Fogla, P., Dagon, D., Lee, W., Skoric, B., 2006b. Towards an information-theoretic framework for analyzing intrusion detection systems. In: Gollmann, D., Meier, J., Sabelfeld, A. (Eds.), Proceedings of the Computer Security – ESORICS 2006. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 527–546.

Hancock, J.C., 1966. Signal Detection Theory. McGraw-Hill.

Hang, J., Han, K., Chen, H., Li, Y., 2020. Ensemble adversarial black-box attacks against deep learning systems. Pattern Recognit. 101, 107184. doi:10.1016/j.patcog.2019.107184.

Hartl, A., Bachl, M., Fabini, J., Zseby, T., 2020. Explainability and adversarial robustness for RNNs. In: Proceedings of the IEEE Sixth International Conference on Big Data Computing Service and Applications (BigDataService), pp. 148–156. doi:10.1109/BigDataService49289.2020.00030.

Hashemi, M.J., Cusack, G., Keller, E., 2019. Towards evaluation of NIDSs in adversarial setting. In: Proceedings of the 3rd ACM CoNEXT Workshop on Big DAta, Machine Learning and Artificial Intelligence for Data Communication Networks. Association for Computing Machinery, New York, NY, USA, p. 1421. doi:10.1145/3359992.3366642.

Hoballah, I., Varshney, P., 1989. Distributed Bayesian signal detection. IEEE Trans. Inf. Theory 35 (5), 995–1000. doi:10.1109/18.42208.

Hogan, R.J., Ferro, C.A.T., Jolliffe, I.T., Stephenson, D.B., 2010. Equitability revisited: why the "equitable threat score" is not equitable. Weather Forecast. 25 (2), 710–726. doi:10.1175/2009WAF2222350.1.

Jiang, Z., Zhao, J., Li, X., Han, J., Xi, W., 2013. Rejecting the attack: source authentication for Wi-Fi management frames using CSI information. In: Proceedings of the IEEE INFOCOM, pp. 2544–2552. doi:10.1109/INFCOM.2013.6567061.

Kingma, D. P., Ba, J., 2015. Adam: a method for stochastic optimization. 1412.6980.
Lee, W., Stolfo, S.J., 2000. A framework for constructing features and models for intrusion detection systems. ACM Trans. Inf. Syst. Secur. 3 (4), 227261. doi:10.1145/382912.382914.
Massey, J. L., 1998. Applied digital information theory I. Lecture Notes, ETH Zurich.
Mell, P., Hu, V., Lippmann, R., Haines, J., Zissman, M., 2003. An overview of issues in testing intrusion detection systems. In: NIST Interagency/Internal Report. National Institute of Standards and Technology, p. 121.
Meng, Y., 2012. Measuring intelligent false alarm reduction using an ROC curve-based approach in network intrusion detection. In: Proceedings of the IEEE International Conference on Computational Intelligence for Measurement Systems and Applications (CIMSA). IEEE, pp. 108–113.
Meng, Y., Kwok, L., 2013. Towards an information-theoretic approach for measuring intelligent false alarm reduction in intrusion detection. In: Proceedings of the 12th IEEE International Conference on Trust, Security and Privacy in Computing and Communications, pp. 241–248. doi:10.1109/TrustCom.2013.33.
Milenkoski, A., Vieira, M., Kounev, S., Avritzer, A., Payne, B.D., 2015. Evaluating computer intrusion detection systems: a survey of common practices. ACM Comput. Surv. 48 (1). doi:10.1145/2808691.
Nasr, K., El Kalam, A.A., 2014. A novel metric for the evaluation of IDSs effectiveness. In: Cuppens-Boulahia, N., Cuppens, F., Jajodia, S., Abou El Kalam, A., Sans, T. (Eds.), ICT Systems Security and Privacy Protection. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 220–233.
Nasr, K., Kalam, A.A.-E., Fraboul, C., 2012. Performance analysis of wireless intrusion detection systems. In: Xiang, Y., Pathan, M., Tao, X., Wang, H. (Eds.), Internet and Distributed Computing Systems. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 238–252.
Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z.B., Swami, A., 2017. Practical black-box attacks against machine learning. In: Proceedings of the ACM on Asia Conference on Computer and Communications Security. Association for Computing Machinery, New York, NY, USA, p. 506519. doi:10.1145/3052973.3053009.
Papernot, N., McDaniel, P., Jha, S., Fredrikson, M., Celik, Z.B., Swami, A., 2016. The limitations of deep learning in adversarial settings. In: Proceedings of the IEEE European Symposium on Security and Privacy (EuroS P), pp. 372–387. doi:10.1109/EuroSP.2016.36.
Paszke, A., Gross, S., Chintala, S., Chanan, G., Yang, E., DeVito, Z., Lin, Z., Desmaison, A., Antiga, L., Lerer, A., 2017. Automatic differentiation in PyTorch. In: Proceedings of the 31st Conference on Neural Information Processing Systems (NIPS 2017).
Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E., 2011. Scikit-learn: machine learning in Python. J. Mach. Learn. Res. 12, 2825–2830.
Puketza, N.J., Zhang, K., Chung, M., Mukherjee, B., Olsson, R.A., 1996. A methodology for testing intrusion detection systems. IEEE Trans. Softw. Eng. 22 (10), 719–729. doi:10.1109/32.544350.
Sharafaldin, I., Lashkari, A.H., Ghorbani, A.A., 2018. Toward generating a new intrusion detection dataset and intrusion traffic characterization. In: Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP), pp. 108–116.
Sheatsley, R., Papernot, N., Weisman, M., Verma, G., McDaniel, P., 2020. Adversarial examples in constrained domains. 2011.01183.
Yan, W., Hylamia, S., Voigt, T., Rohner, C., 2020. Phy-ids: a physical-layer spoofing attack detection system for wearable devices. In: Proceedings of the 6th ACM Workshop on Wearable Systems and Applications. Association for Computing Machinery, New York, NY, USA, p. 16. doi:10.1145/3396870.3400010.
Zhang, Y., Liu, P., Liu, Y., Li, A., Du, C., Fan, D., 2013. Attacking pattern matching algorithms based on the gap between average-case and worst-case complexity. J. Adv. Comput. Netw. 1 Number 3, 228–233.
Tramér, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., McDaniel, P., 2020. Ensemble adversarial training: attacks and defenses. 1705.07204.

Tobias Mages is currently pursuing a Ph.D. degree in the Department of Information Technology, Division of Computer Systems at the Uppsala University. He received a bachelor's degree in Electrical Engineering/Communications Engineering from the Baden-Wuerttemberg Cooperative State University (DHBW), Germany in 2017 and a master's degree in Embedded Systems from the Uppsala University, Sweden in 2019. His research interests are in network-based intrusion detection systems (NIDS) and internet-of-things (IoT) security.

Magnus Almgren is an Associate professor in cyber-physical systems at Chalmers investigating security properties of systems with a large societal impact. Dr. Almgren has been a Fulbright Scholar and holds an MS in Engineering Physics from Uppsala University, an MS in Computer Science with distinction in research from Stanford University, and a Ph.D. in Computer Science from Chalmers University of Technology. His expertise is in application-based intrusion detection systems (IDS) and reasoning about conflicting information from several detectors in a larger system.

Christian Rohner is Professor in computer systems at Uppsala University working with wireless communication and security. Dr. Rohner holds an M.Sc. in Electrical Engineering and Ph.D. in Computer Science from ETH Zürich. His research interests are related to resource constrained wireless systems, in particular ultra-low power communication and wireless security.