Poster: Ensemble Learning for Industrial Intrusion Detection

Dominik Kus[*], Konrad Wolsing[†,*], Jan Pennekamp[*], Eric Wagner[†,*],
Martin Henze[‡,†], and Klaus Wehrle[*]

[*] *Communication and Distributed Systems*, RWTH Aachen University, Germany · {lastname}@comsys.rwth-aachen.de
[†] *Cyber Analysis & Defense*, Fraunhofer FKIE, Germany · {firstname.lastname}@fkie.fraunhofer.de
[‡] *Security and Privacy in Industrial Cooperation*, RWTH Aachen University, Germany · henze@cs.rwth-aachen.de

ABSTRACT

Industrial intrusion detection promises to protect networked industrial control systems by monitoring them and raising an alarm in case of suspicious behavior. Many *monolithic* intrusion detection systems are proposed in literature. These detectors are often specialized and, thus, work particularly well on certain types of attacks or monitor different parts of the system, *e.g.*, the network or the physical process. *Combining* multiple such systems promises to leverage their joint strengths, allowing the detection of a wider range of attacks due to their diverse specializations and reducing false positives. We study this concept's feasibility with initial results of various methods to combine detectors.

1 MOTIVATION

Industrial Control Systems (ICSs) are increasingly connected to the Internet, and thereby, they are exposed to sophisticated cyberattacks [3]. Such attacks can cause severe damage [2], making ICSs a valuable target, particularly for state-level actors, as the Ukrainian Power Grid attack [7] prominently proved. Consequentially, there is a strong demand to secure ICSs and protect critical infrastructure.

Industrial Intrusion Detection Systems (IIDSs) provide an additional layer of security by monitoring an ICS's largely repetitive physical processes and network communication patterns and notifying the operators in case of a suspected attack. To this end, *state*-based IIDSs monitor physical process parameters for manipulations while *network*-based IIDSs detect anomalies in the system's network traffic, such as Denial of Service attacks [8, 17].

Existing work focuses on devising effective *monolithic* detectors, which are often specialized and, thus, detect certain types of attacks particularly well. *Combining* multiple such systems into an IIDS ensemble could leverage their strengths and specializations to enable the detection of a wider range of attacks, reduce false positives and improve the detection rate. While ensemble learning, such as weighted voting, has been proposed in general [1, 4, 5, 10, 15, 18], applying its methods and concepts to IIDSs remains unexplored.

2 ENSEMBLE LEARNING FOR IIDSS

In contrast to monolithic approaches, IIDS ensembles offer a multitude of advantages. First, ensembles promise to avoid the lock-in to a single approach and thereby reduce the risk of missing attacks since a combination of IIDSs designed for different types of attacks can complement each other. Next, fusing alerts from several similar IIDSs may enhance their collective certainty, *i.e.*, leading to fewer false alarms [6], which is crucial since the majority of an ICS's traffic occurs under normal operation without the presence of attacks. Furthermore, assembling an IIDS out of a wider variety of monolithic classifiers could result in a system that is overall more flexible, *i.e.*, leveraging signatures to (re-)identify variations of known attacks and anomaly detection to detect unknown ones.

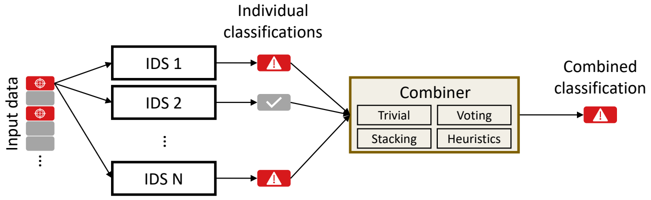

Figure 1: An ensemble approach promises to combine the capabilities of different types of IIDSs, which allows the system to detect a wider range of attacks or reduce false positives.

In this work, we assess the feasibility of various ensembles methods, which we call *combiners*, to leverage the combined detection capabilities of a multitude of IIDSs (*cf.* Fig. 1). We begin with a collection of monolithic IIDSs from related work, which each emit an independent prediction of whether the ICS is in a benign or anomalous state. The combiner then takes the IIDSs' outputs and is in charge of deriving a final classification, *e.g.*, by a majority vote.

3 PRELIMINARY EVALUATION

Given the promising properties of IIDS ensembles (cf. Sec. 2), we aim to explore which combiners (*i.e.*, ensemble methods) can bring a real-world benefit to ICS security. To this end, we experiment with different combiners (Sec. 3.1) and examine to which extent they can hold up to the theoretically achievable performance (Sec. 3.2).

We apply seven machine learning-based classifiers proposed in literature based on Random Forest (RF), Support Vector Machine (SVM), Bidirectional Long Short Term Memory (BLSTM) [9], Decision Trees, Extra Trees (ET), Isolation Forest [16], and Naïve Bayes [14], provided by the open-source IPAL framework [17], and consider two datasets: a power system [12] and a gas pipeline [11]. We split each dataset into a train set for the classifiers (40 %), a second train set for the combiner (40 %), and a test set for evaluation (20 %). The combiner train set is needed as the IIDSs' performance on their train set is not indicative of their behavior on new data.

In the following, we compare various IIDS combiners against the best monolithic IIDS for each dataset (Extra Trees for the power system and Random Forest for the gas pipeline dataset).

3.1 Initial Combiner Results

We start with simple rules to combine multiple IIDSs: *all* (all IIDSs must emit an alert), *any* (at least one IIDS emits an alert), and a *majority vote* [18]. For the IIDSs and datasets under study, such rules perform worse than the best individual classifier (*cf.* Tab. 1),

Poster to be presented at *Annual Computer Security Applications Conference (ACSAC '22), December 05–09, 2022, Austin, Texas, USA*
2022. This is the author's version of the work. It is posted here for your personal use.

Table 1: Exploring different combining methods reveals a slight improvement on the gas pipeline dataset and a roughly matching performance on the power system dataset. The achievable improvement (upper bound) is, however, marginal and limited by the classifier selection.

Classifier [%]	Power System Dataset [12]			Gas Pipeline Dataset [11]		
	Acc.	F1	FPR	Acc.	F1	FPR
Best Individual	80.72	87.12	47.32	98.74	97.04	0.29
All	−47.84	−76.68	−45.95	−20.46	−96.62	−0.29
Any	−10.06	−4.32	+52.57	−08.18	−15.04	+11.44
Majority Vote	−0.69	−0.34	+3.32	−0.60	−1.47	−0.11
Best Weighted Vote	−0.52	+0.10	+9.68	+0.15	+0.37	+0.07
LR	−0.17	−0.04	+1.65	+0.12	+0.31	+0.13
SVM	−0.18	−0.13	+0.02	+0.17	+0.43	+0.20
Upper Bound	+0.48	+0.22	−2.89	+0.21	+0.53	+0.20

but the majority vote falls behind by only 0.34 %/1.47 % in the F1 score on the power system and gas pipeline dataset. Yet, it improves the FPR on the gas pipeline dataset by 0.11 %. Unsurprisingly, the all combiner achieves the lowest FPR of only 1.37 %/0.0 %, showing that a tradeoff can be made between detection accuracy and FPR.

As an improvement to simple rules, we assign weights w_i to each IIDS's alerts and raise an alarm if the sum of weights surpasses a threshold t. We choose the weights on a best effort basis. For the gas pipeline dataset, $w_{\text{BLSTM}} = w_{\text{ET}} = 1$, $w_{\text{RF}} = 2$, $t = 2$, and $w_c = 0$ for all other c yielded the best performance. This configuration outperforms the baseline by 0.37 % in the F1 score, only falling short by 0.07 % in FPR. On the power system dataset, $w_{\text{ET}} = w_{\text{RF}} = 1$, $t = 1$, and $w_c = 0$ for all other c outperforms the baseline merely by 0.1 % in the F1 score but falls short in accuracy and FPR.

While manually chosen weights may be sub-optimal, finding a good combiner constitutes a machine learning problem. The goal is to learn an optimal mapping between the base IIDSs' outputs and the expected classification result, also known as *stacking* [13]. To this end, we leverage Logistic Regression (LR) and SVMs, which outperform the best individual classifier on the gas pipeline dataset in terms of F1 score by 0.31 % and 0.43 % for LR and SVM, respectively.

In summary, we found that weighted voting can perform well but requires manually chosen weights, while stacking yields similar results without manual adjustments. Of all our tested combiners, SVM performs best in terms of accuracy and F1 score. While the results look promising on the gas pipeline dataset, the power system dataset's best classifier could not be surpassed by any of our tested combiners, leaving us with the question of how big the available headroom for improvement actually is.

3.2 A Practical Upper Bound

Consequently, we want to estimate an upper bound on the achievable performance by any combiner. To this end, we leverage a heuristic similar to the Behavior Knowledge Space Method [18]. It creates a mapping to the expected label for each of the $2^7 = 128$ possible output combinations of the seven IIDSs. The heuristic maps to malicious if the majority of outputs during training are malicious or benign otherwise. This approach guarantees the lowest possible amount of misclassifications, thus maximizing accuracy.

From Tab. 1, we observe very little headroom in terms of accuracy compared to the best individual classifier on both datasets, with 0.48 % for the power system and 0.21 % for the gas pipeline dataset. On the latter, the SVM combiner leaves a gap of only 0.04 %. Consequently, practical ensemble learning techniques can reach nearly optimal combination performance, but they heavily rely on a diverse set of input IIDS to substantially improve the performance.

4 CONCLUSION AND FUTURE WORK

IIDS ensembles promise to boost the detection performance and combine the strengths of different approaches, *e.g.*, state- and network-based detectors. While showing that ensemble methods are applicable in general, our first results are mixed, only marginally improving upon the best individual classifier on the gas pipeline dataset and roughly matching it on the power system dataset.

We assume that these results are mainly caused by a lacking diversity in our classifier selection, which consists only of state-based supervised machine learning classifiers thus far. This assumption is supported by our upper bound, which shows that the best improvement in accuracy is less than 0.5 % on both datasets. Integrating different types of classifiers, however, proves non-trivial. To tackle this, we plan to independently evaluate ensemble methods for network-based IIDSs, which pose a different set of challenges, and then integrate both types of IIDSs to unlock additional potential.

Another option to improve ensemble results, especially in ambiguous cases, is providing the combiner with more data, *e.g.*, by utilizing the classifiers' internal confidence values, which express how confident the IIDS is in its prediction, in addition to their binary predictions. With a more diverse set of classifiers or in settings without well-performing individual IIDSs, we believe that ensemble learning can deliver on its promise to widen the range of detectable attacks while reducing the number of false positives.

REFERENCES

[1] Abdulla Amin Aburomman and Mamun Bin Ibne Reaz. 2017. A survey of intrusion detection systems based on ensemble and hybrid classifiers. *Comput. Secur.* 65.

[2] Tejasvi Alladi et al. 2020. Industrial control systems: Cyberattack trends and countermeasures. *Comput. Commun.* 155.

[3] Deval Bhamare et al. 2020. Cybersecurity for industrial control systems: A survey. *Comput. Secur.* 89.

[4] Xiayang Chen et al. 2018. Ensemble Learning Methods for Power System Cyber-Attack Detection. In *IEEE ICCCBDA*.

[5] Pandit Byomakesha Dash et al. 2020. Model based IoT security framework using multiclass adaptive boosting with SMOTE. *Secur. Priv.* 3, 5.

[6] Dominik Kus et al. 2022. A False Sense of Security? Revisiting the State of Machine Learning-Based Industrial Intrusion Detection. In *ACM CPSS*.

[7] Robert M. Lee et al. 2016. *Analysis of the Cyber Attack on the Ukrainian Power Grid.* Defense Use Case 5. E-ISAC and SANS.

[8] Hung-Jen Liao et al. 2013. Intrusion detection system: A comprehensive review. *J. Netw. Comput. Appl.* 36, 1.

[9] Rocio Lopez Perez et al. 2018. Machine Learning for Reliable Network Attack Detection in SCADA Systems. In *IEEE TrustCom*.

[10] Joris Lueckenga et al. 2016. Weighted Vote Algorithm Combination Technique for Anomaly Based Smart Grid Intrusion Detection Systems. In *IJCNN*.

[11] Thomas H. Morris et al. 2015. Industrial control system simulation and data logging for intrusion detection system research. In *SCSS*.

[12] Shengyi Pan et al. 2015. Developing a Hybrid Intrusion Detection System Using Data Mining for Power Systems. *IEEE Trans. Smart Grid* 6, 6.

[13] Omer Sagi and Lior Rokach. 2018. Ensemble learning: A survey. *WIREs Data Min. Knowl. Discov.* 8, 4.

[14] Syed Noorulhassan Shirazi et al. 2016. Evaluation of Anomaly Detection techniques for SCADA communication resilience. In *RWS*.

[15] Bayu Adhi Tama and Sunghoon Lim. 2021. Ensemble learning for intrusion detection systems: A systematic mapping study and cross-benchmark evaluation. *Comput. Sci. Rev.* 39.

[16] Herman Wijaya et al. 2020. Domain-Based Fuzzing for Supervised Learning of Anomaly Detection in Cyber-Physical Systems. In *IEEE/ACM ICSEW*.

[17] Konrad Wolsing et al. 2022. IPAL: Breaking up Silos of Protocol-dependent and Domain-specific Industrial Intrusion Detection Systems. In *RAID*.

[18] Zhi-Hua Zhou. 2012. *Ensemble Methods: Foundations and Algorithms.* CRC Press.