

# Evaluation of Machine Learning-based Anomaly Detection Algorithms on an Industrial *Modbus/TCP* Data Set

Simon Duque Anton, Suneetha Kanoor, Daniel Fraunholz, and Hans Dieter Schotten

Intelligent Networks Research Group

German Research Center for Artificial Intelligence

Kaiserslautern, Germany

{simon.duque\_anton,suneetha.kanoor,daniel.fraunholz,hans\_dieter.schotten}@dfki.de

## ABSTRACT

In the context of the *Industrial Internet of Things*, communication technology, originally used in home and office environments, is introduced into industrial applications. Commercial off-the-shelf products, as well as unified and well-established communication protocols make this technology easy to integrate and use. Furthermore, productivity is increased in comparison to classic industrial control by making systems easier to manage, set up and configure. Unfortunately, most attack surfaces of home and office environments are introduced into industrial applications as well, which usually have very few security mechanisms in place. Over the last years, several technologies tackling that issue have been researched. In this work, machine learning-based anomaly detection algorithms are employed to find malicious traffic in a synthetically generated data set of *Modbus/TCP* communication of a fictitious industrial scenario. The applied algorithms are *Support Vector Machine (SVM)*, *Random Forest*, *k-nearest neighbour* and *k-means clustering*. Due to the synthetic data set, supervised learning is possible. *Support Vector Machine* and *k-nearest neighbour* perform well with different data sets, while *k-nearest neighbour* and *k-means clustering* do not perform satisfactorily.

This is a preprint of a work published in the Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES 2018). Please cite as follows:

S. D. Duque Anton, S. Kanoor, D. Fraunholz, and H. D. Schotten: "Evaluation of Machine Learning-based Anomaly Detection Algorithms on an Industrial *Modbus/TCP* Data Set." In: Proceedings of the 13th International Conference on Availability, Reliability and Security (ARES 2018), ACM, 2018, pp. 41:1–41:9.

## CCS CONCEPTS

• **Security and privacy** → *Intrusion detection systems*;

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from [permissions@acm.org](mailto:permissions@acm.org).

ARES 2018, August 27–30, 2018, Hamburg, Germany

© 2018 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-1-4503-6448-5/18/08.

<https://doi.org/10.1145/3230833.3232818>

## KEYWORDS

Modbus, Machine Learning, Anomaly Detection, Industrial, IT-Security

## ACM Reference Format:

Simon Duque Anton, Suneetha Kanoor, Daniel Fraunholz, and Hans Dieter Schotten. 2018. Evaluation of Machine Learning-based Anomaly Detection Algorithms on an Industrial *Modbus/TCP* Data Set. In *ARES 2018: International Conference on Availability, Reliability and Security, August 27–30, 2018, Hamburg, Germany*. ACM, New York, NY, USA, 9 pages. <https://doi.org/10.1145/3230833.3232818>

## 1 INTRODUCTION

Since the appearance of industrial control in the 1970's, industry has been looking for ways to improve production. At first, hard-wired sensors and actuators were employed, followed by so-called Supervisory Control And Data Acquisition (SCADA) systems in the 1980's and '90's. With the emerging computational and communication technology, the automation pyramid, as depicted in figure 1, arose. It categorised the industrial networks according to

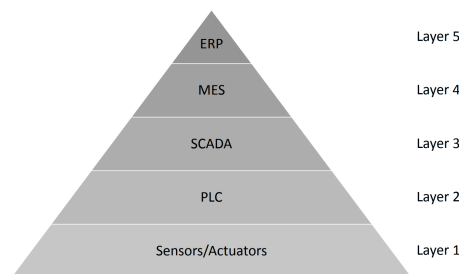


Figure 1: Automation Pyramid

their function: Resource and production planning was done on the two topmost layers, four and five. SCADA systems are located at the third layer. Programmable Logic Controller (PLC)s can be found on the second layer, while sensors and actuators are placed on the first layer. This has been possible mainly due to Commercial Off The Shelf (COTS) products that are interchangeable and configurable. Many of the newly introduced network protocols either base on the Ethernet-protocol on layer 2 of the Open System Interconnection (OSI)-model, such as *EtherCAT* [14] and *Modbus* [29, 30], or even on the Transmission Control Protocol (TCP)/Internet Protocol (IP)-stack on layers 3 and 4 of said OSI-model. An abundance of proprietary and open-source communication protocols, tailored

to the needs of industrial applications, was developed. Prominent, TCP/IP-based examples are *Modbus/TCP* [28, 29], *ProfiNET* [35] and *OPC Unified Architecture (UA)* [34]. After integrating communication protocols, based on the OSI-model, and introducing re-programmable industrial computers, so-called PLCs, industrial and automation networks are being opened to insecure networks. The arising Industrial Internet of Things (IIoT) requires interconnectivity of networks, accessibility and availability of resources, also outside of trust boundaries. A common assumption of early SCADA-implementations was that networks were physically separated from public networks [21], breaking the trust boundaries creates a multitude of novel threats and attack vectors [11, 21, 52]. Attackers have identified industrial networks as valuable targets. Widespread bot-nets provide easy opportunities to probe and capture unprotected Internet of Things (IoT)-devices [16]. This novel threat landscape necessitates new approaches for intrusion detection and attack prevention. Machine learning technologies, and methods of artificial intelligence have proven that they provide vast capacities for solving problems that were hard to solve otherwise.

The remainder of this work is structured as follows: Related work in artificial intelligence and machine learning for intrusion detection in industrial networks is presented in section 2. After that, the employed data set is presented in section 3. First, *Modbus/TCP* is presented as a protocol. Second, the data set is evaluated, and third, features are derived. The algorithms for anomaly detection and their application on the data set is described in section 4. Results are discussed in section 5. Finally, a conclusion is drawn in section 6.

## 2 RELATED WORK

Intrusion detection in office Information Technology (IT) environments is a well-researched and well-established area. Tools, such as *Bro* [36] and *Snort* [43], are commonly used and maintained by a widespread user base. They allow for easy integration of custom rules and make efficient firewalls and systems for detecting attackers and intrusions. The same holds for data sets of host and network traffic. There are numerous data sets to train and test Intrusion Detection System (IDS) appliances and machine learning methods, one of the most famous being the ‘99 KDD Cup data set [48]. As this work focuses on industrial applications of anomaly detection, IDSs for office applications, as well as data sets with home- and office-based network traffic are not considered here.

One of the most important aspects in the development of novel intrusion detection approaches is a sound data set to test the system and verify the findings. As mentioned, there is an abundance of such data sets for home and office network traffic, while industrial network traffic is still relatively rare. One recent data set is presented by *Lemay and Fernandez* [24]. They propose an architecture for a traffic simulation environment, based on commonly used *Modbus/TCP* tools and sandbox environments. They also published a data set into which malicious traffic has been introduced. This data set was analysed in this work in order to evaluate the effectiveness of machine learning for industrial intrusion detection. Other than that, *Wang et al.* propose a simulation environment for SCADA security analysis [50]. Their framework allows setting up *OPC UA*

components, including sensors and actuators, in a simulation in order to test and verify security solutions. Furthermore, *Siaterlis et al.* propose a testbed for the effects of cyber attacks on Cyber Physical System (CPS) [42]. The testbed is based on an *Emulab* [45] emulation environment and is capable of monitoring the impact of an attack on a production system. *Genge et al.* follow a similar approach [18]. They present an adaptable testbed that is capable of emulating different industrial production scenarios. These scenarios can then be attacked with real malware and the effects can be evaluated. Their testbed is based on *Emulab* [45] as well, with a real-time connection simulator *Simulink* [26]. *Seidl* designed a *Python* [15]-environment that simulates user-defined industrial behaviour called *VirtuaPlant* [41]. This simulation can then be introduced to attacks and malware.

In addition to simulation environments, there are also data sets available in order to train intrusion and anomaly detection algorithms. *Morris and Gao* present several files containing sets of industrial control system traffic [31, 32]. As malicious traffic is introduced into these data, algorithms can be trained to detect traffic of malware.

Apart from the issue of obtaining sound and plausible data, there is an abundance of algorithms for anomaly detection that could be employed in order to detect intrusions. Intrusion detection as a concept, including a formal model, was originally presented by *Denning* in 1987 in the context of the growing influence of computer systems and networks [9]. The applications of anomaly detection mechanisms for network intrusion detection are discussed in several surveys [4, 22]; *Yang et al.* give a brief introduction of these techniques for the domain of SCADA systems [51]. *Meshram and Haas* published a roadmap of machine learning based anomaly detection in industrial networks, containing a simulation environment, as well as a semantic description of content [27]. *Kleinman and Wool* present a model of the *Siemens S7* protocol for intrusion detection and forensics [23]. Critical infrastructures and industrial environments are considered in the work of *Hadziosmanovic et al.* [20]. A framework that detects malicious and undesired actions is presented. Deriving features that can be used to distinguish valid from malicious traffic is the first step in applying an intrusion detection algorithm. *Mantere et al.* look into the derivation of features from IP traffic in an industrial environment [25]. Deterministic properties of industrial control systems, as well as the usability of this feature for anomaly detection in an industrial environment, is researched by *Hadeli et al.* [19].

## 3 DISSECTING THE DATA SET

In this section, the data set is described. First, a general introduction to the *Modbus* protocol is given in subsection 3.1. After that, the data set used in the course of this work, presented by *Lemay and Fernandez* [24], is described in subsection 3.2. Finally, the features that have been extracted and derived are presented in subsection 3.3.

### 3.1 An Introduction to *Modbus*

*Modbus* is a communication protocol for serial communication among PLCs and Remote Terminal Units (RTUs). It has been developed in 1979 by *Schneider-Electric*, formerly known as *Modicon* [40]. It has become a de-facto standard communication protocol for industrial communication [10]. There are several versions of *Modbus* available, the most noteworthy are listed in table 1. In *Modbus/TCP*,

**Table 1: An Overview of the Most Used *Modbus* Versions**

Version	Description
<i>Modbus RTU</i>	Serial communication via RS-232 connector to connect PLCs with RTUs
<i>Modbus ASCII</i>	Same connector as above, but instead of binary coding, ASCII-encoded characters are used
<i>Modbus TCP/IP</i>	Communication based on the TCP/IP protocol stack
<i>Modbus over TCP/IP</i>	Same as above, but including a checksum in the payload, in addition to error correction mechanisms provided by layers 1 to 4 of the OSI model

communication is encapsulated in a TCP/IP packet, as shown in figures 2 and 3. They are transmitted via ethernet, which follows the structure depicted in figure 4. All dark gray fields were employed as features in this work. Most *Modbus/TCP* messages contain com-

mands regarding reading and writing coils or registers. In analogy to analogue control automation, one bit registers are called *coils*. Multi-bit registers are called *registers*. *Modbus* slaves poll their communication and either set the data as new input for their registers or load information into a register for a master to read it. They then respond to the request.

### 3.2 Description of the Data Set

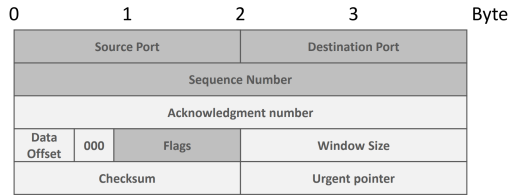
*Lemay and Fernandez* simulated a controller network, consisting of a number of Master Terminal Unit (MTU) and of a number of controllers. The controllers control a simulated physical system with a 12 000 Volt power source, as well as main- and sub-branch cut-off breakers. In this scenario, different data collections have been performed. An exhaustive description can be found in their work [24]. Regular polling and manual operation are part of these data sets, as these actions occur in productive systems in this form. After collection, malicious activities, generated by state of the art penetration testing tools such as *metasploit* [37], are introduced.

This is one of the most prevalent drawbacks of the employed data set: according to *Morris and Gao* there are several different groups of *Modbus*-based attacks [31]. Unfortunately, none of these is introduced into this data set. Instead, attacks that are also common in home and office-based penetration testing are introduced. Unfortunately, this does not mimic a wide range of attacks that could be employed against industrial applications. It does, however, mimic the timing behaviour and the rate of packets per time unit, which is a good distinguishing factor for attacks.

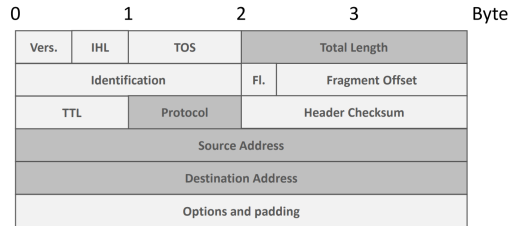
In this work, three data sets, henceforth called *DS1* to *DS3*, were used for testing the algorithms:

- *DS1: Moving\_two\_files\_Modbus\_6RTU*: Regular traffic between one MTU and six RTU during three minute interval, 3 319 packets captured, contains 75 malicious instances
- *DS2: Send\_a\_fake\_command\_Modbus\_6RTU\_with\_operate*: Regular traffic between one MTU and six RTU during 10 minute interval, 11 166 packets captured, contains 10 malicious instances
- *DS3*: A combination of eight data sets, four of which do and four of which do not contain malicious activity, 365 906 packets overall, contains 206 malicious instances

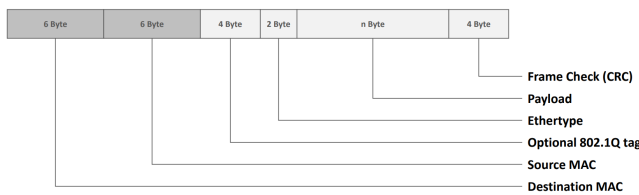
*DS3* addresses a common problem in real-world intrusion detection with machine learning: In order to train the algorithm, a normal condition of the system has to be derived, deviations from which have to be recognized as anomalies. A common practice is to monitor the behaviour of the productive system for a certain time under the assumptions that it does not contain malicious traffic. There are two issues, however. First, in productive systems, you can never be sure that there is no malicious traffic. It is just highly unlikely. Second, the recognition of anomalies based on normal behaviour can be difficult, as the user usually does not know the characteristics that have most impact on the algorithm. The effect of these limitations are evaluated in this work by mixing different kinds of traffic, even traffic with no malicious content. Due to the synthetic nature of this traffic, one can be sure that it is non-malicious.



**Figure 2: TCP Frame Structure**



**Figure 3: IP Frame Structure**



**Figure 4: Ethernet Frame Structure**

### 3.3 Feature Extraction

The first step in anomaly detection and data mining is the determination of relevant features. These features can be used to describe the data instances with respect to a given goal; the goal in the given case is to determine instances that differ significantly from the common, productive behaviour. Hence, features that are suited to describe the normal behaviour of the system are needed. In general, there are two different kinds of features: Basic and derived features. Basic features are already present within the data. In the given case, they are contained within the protocol headers. Network traffic, for example, contains source and destination addresses, lengths, time stamps and other features. An exhaustive list containing the 14 basic features of this data set can be found in table 2. Two features are derived from the ethernet header as shown in figure 4, four features each are obtained from TCP and IP header, as shown in figures 2 and 3, two features from User Datagram Protocol (UDP) headers respectively and two features from the capturing tool, namely arrival time and information about broken packets. Derived features

**Table 2: The Basic Features Considered in this Work**

Feature	Description
<i>frame.number</i>	Sequential number of packet
<i>frame.time</i>	Arrival time of packet with millisecond accuracy
<i>eth.src</i>	Ethernet source address (MAC)
<i>eth.dst</i>	Ethernet destination address (MAC)
<i>ip.src</i>	IP source address
<i>ip.dst</i>	IP destination address
<i>ip.proto</i>	Transport Layer protocol
<i>frame.len</i>	Length of IP-packet in bytes
<i>tcp.flags</i>	Control bits of TCP-packet
<i>tcp.srcport</i>	Port number of source in TCP connection
<i>tcp.dstport</i>	Port number of destination TCP connection
<i>udp.srcport</i>	Port number of source UDP connection
<i>udp.dstport</i>	Port number of destination UDP connection
<i>tcp.analysis.lost_segment</i>	A label set if there is a lost segment

result from the combination of basic features and can often only be derived from sequences of packets, e.g. the number of packets per time unit. Given the time stamp and the number of bytes of each packet, for example, the transmitted amount of bytes per second can be calculated. A list of nine derived features generated from this data set can be found in table 3. The impact of each feature on

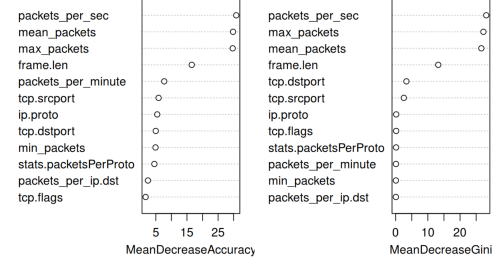
**Table 3: The Derived Features Considered in this Work**

Feature	Description
<i>frame.time.min</i>	Time of frame in minutes
<i>packets_per_minute</i>	Number of packets per minute
<i>frame.time.sec</i>	Time of frame in seconds
<i>packets_per_sec</i>	Number of packets per second
<i>packets_per_ip.dst</i>	Number of packets per unique destination-IP
<i>stats.packets_per_proto</i>	Number of packets per protocol
<i>max_packets</i>	Maximum number of packets per second
<i>min_packets</i>	Minimum number of packets per second
<i>mean_packets</i>	Mean number of packets per second

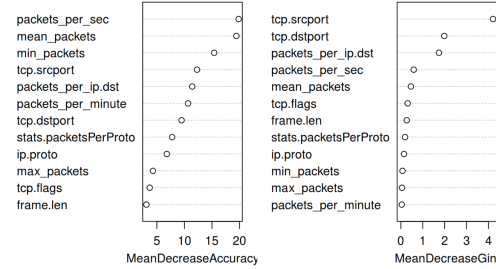
the prediction can be calculated. In order to do so, the decrease of accuracy of the prediction is evaluated. The higher the decrease,

the more important the feature. Another metric for the importance of a feature is the decrease in Gini index. The Gini index describes the pureness of a data set, split according to a given feature [38]. The higher the decrease in Gini index, the more a feature is suited to split a data set into anomalous and non-anomalous.

The *packets\_per\_second*, *mean\_packets* and *max\_packets* are the features with the highest impact on the result for data sets *DS1* and *DS2* as shown in figures 5 and 6. The fact that all of them are derived features underlines the importance of feature engineering. Due

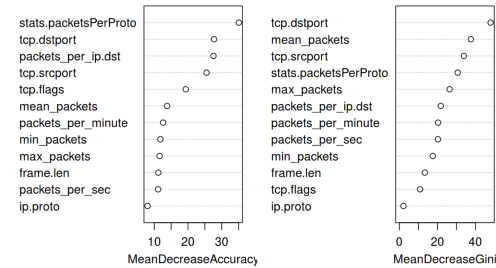


**Figure 5: Importance of Features for svm in DS1**



**Figure 6: Importance of Features for svm in DS2**

to the characteristic of *DS3*, consisting of different kinds of traffic, another feature importance occurs, as depicted in figure 7. The TCP



**Figure 7: Importance of Features for svm in DS3**

destination port are of importance in differentiating, as well as the number of packets per protocol and per destination IP. Furthermore, the TCP source port distribution and the mean number of packets per second are important. In this scenario, some basic features are of high importance for the anomaly detection. A sound understanding of the scenario and application area therefore is of the essence.

#### 4 ANOMALY DETECTION IN MODBUS DATA

In this section, the application of four different machine learning algorithms, namely *Support Vector Machines (SVM)*, *Random Forrest*, *k-nearest neighbour* and *k-means clustering*, is described. Those algorithms are used to find outliers in the three data sets *DS1*, *DS2* and *DS3*, described in subsection 3.2, using the features presented in subsection 3.3. At first, the data sets are split into 70% and 30%, as well as 80% and 20% respectively for cross-validation. The split values used depend on the quality of the cross-validation, the one providing better results is chosen. In this work, an 80%/20% has only been chosen for the *k-nearest neighbour* as described in subsection 4.4. The larger part is used to train the algorithm. Due to the labels, the prediction of an algorithm can be compared with the label in order to determine whether the prediction was correct or not. After training, the remaining part is used for testing. In this phase, the algorithm isn't adjusted anymore. Still, the predictions are compared to the labels in order to determine metrics that describe the quality of an outlier detection algorithm.

Outlier detection can be seen as a binary classifier: An instance is either normal or anomalous. There are several metrics available to determine the performance of a binary classifier. For intrusion detection in industrial, but also in home and office networks, not only the number of detected attacks is relevant. Due to the high amount of traffic, false positives have severe effects. For one, they need a lot of time to investigate. Furthermore, they can, on a psychological level, have administrators become careless in cases of alarms as they expect them to be false positives, the so called *alarm fatigue* [5]. Finally, the amount of normal traffic in networks usually outnumbers the amount of malicious traffic by magnitudes. That means wrongly classifying 0.1% of malicious and of normal traffic still results in vastly different numbers of alarms. In this work, we used two metrics to describe the performance of the algorithms: The *accuracy* [33], as well as the *f-measure* [49]. The *f-measure*, or *F1-score*, is calculated as described in equation 1.

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}} \quad (1)$$

$$\text{precision} = \frac{t_p}{t_p + f_p} \quad (2)$$

$$\text{recall} = \frac{t_p}{t_p + f_n} \quad (3)$$

$t$  stands for a correct classification of the algorithm,  $f$  for an incorrect one. An index  $p$  indicates that the algorithm classified it as positive, an index of  $n$  indicates a classification as negative. The *F1-score* provides information about the relation of *precision* and *recall*, as defined in equations 2 and 3. *Precision* and *recall* describe the relation of all true positive classifications to all that have been classified as positive, respectively to all events that are positive. If both values are perfect, the *F1-score* amounts to one; at worst, it reaches 0. The *accuracy* is calculated according to equation 4.

$$\text{accuracy} = \frac{t_p + t_n}{t_p + f_p + t_n + f_n} \quad (4)$$

*Accuracy* gives information about the relation of correct classifications in relation to all classifications. First, a naive approach

to find outliers is described in subsection 4.1. In subsections 4.2 to 4.5, the algorithms are applied to the three data sets *DS1*, *DS2* and *DS3* and their performance is evaluated with the given metrics. The results are then discussed in section 5.

##### 4.1 Naive Approach

In some data sets, exploratory data analysis can lead to the discovery of singular or a group of features that can be used to distinguish between normal and anomalous data. In *DS1*, there are three features of the derived features as explained in table 3 capable of splitting the data set perfectly. These features with the according values are listed in table 4. This makes the application of machine learning

**Table 4: Features Capable of Perfectly Splitting *DS1***

Feature	Norm. Values [Packets/s]	Anom. Values [Packets/s]
<i>packets_per_sec</i>	162, 164	3-9, 41
<i>max_packets</i>	72	208, 235, 401
<i>mean_packets</i>	58,28, 58,48	105-150

algorithms obsolete, all of the applied algorithms have to compete against a perfect score. It is noteworthy, however, that these are derived and not basic features. So at least a thorough understanding and sound feature engineering are necessary in order to be able to make sense of the data.

For *DS2* and *DS3*, no such features exist. *DS2* is too large with too few anomalous instances, so that each feature of an anomalous event takes the same value on at least one other normal event. *DS3* is even more difficult, as several data traces are mixed. This leads to more heterogeneous feature distributions, making it impossible to classify it by exploratory data analysis.

##### 4.2 Support Vector Machines

*SVM* were first introduced by *Boser et al.* in 1992 [6]. The idea is to create a divider between two groups in such a way that each instance has the most possible distance from the divider. This is called a *large margin classifier*. In *SVM*, data points are described by tuples as shown in equation 5 [8].

$$(x_i, y_i), i = 1, \dots, m, y \in \{-1, 1\} \quad (5)$$

$x$  is a vector describing a data point in an  $n$ -dimensional feature space.  $y$  describes the attribution to one of the two classes.  $m$  is the number of data points. After training data, the attribution is performed by the signum-function, as shown in equation 6.  $w$  is the normal vector of the separator hyperplane,  $b$  is the offset from the hyperplane.

$$y_i = \text{sgn}(w \cdot x_i - b) \quad (6)$$

Generally, when applying *SVM*, there are two different cases: Either the set of instances can or can not be divided by a linear geometric figure. If no linear division of the set of instances is possible, the so-called *kernel trick* is applied [8]. In using the *kernel trick*, the input space is mapped non-linearly into a higher dimensional feature space, where the algorithm can create a linear divider. In this

work, the *e1071*-library [46] of the *R* programming language [44] has been used with a linear kernel.

*DS1*: *SVM* performs exceedingly well with this data set. The relation of true and predicted labels can be found in table 5. *SVM*

**Table 5: Predictions and Correct Labels of *DS1* by Using *SVM***

Label\Prediction	Normal	Anomalous
Normal	1 097	0
Anomalous	0	22

is capable of predicting each instance of the test data set correctly, leading to an *accuracy*, as well as *F1-score* of 1, as shown in line 1 of table 6.

**Table 6: Accuracy and *F1-score* of *SVM***

Dataset	Accuracy	<i>F1-score</i>
<i>DS1</i>	1,0	1,0
<i>DS2</i>	1,0	1,0
<i>DS3</i>	0,999 936	0,999 968

*DS2*: *SVM* performs exceedingly well with this data set as well. The relation of true and predicted labels can be found in table 8. *SVM* is capable of predicting each instance of the test data set

**Table 7: Predictions and Correct Labels of *DS2* by Using *SVM***

Label\Prediction	Normal	Anomalous
Normal	3 364	0
Anomalous	0	3

correctly, leading to an *accuracy*, as well as *F1-score* of 1, as shown in line 2 of table 6.

*DS3*: For this data set, *SVM* still performs relatively well. The relation of true and predicted labels can be found in table 8. *SVM*

**Table 8: Predictions and Correct Labels of *DS2* by Using *SVM***

Label\Prediction	Normal	Anomalous
Normal	109 702	4
Anomalous	3	63

is capable of predicting most instances of the data set correctly, indicated by *accuracy* and *F1-score* as shown in line 3 of table 6.

### 4.3 Random Forrest

A collection of *Decision Trees* is called a *Random Forest* [7]. It consists of a *root* node, internal nodes, so-called *split* nodes and *leaf* nodes. Each *leaf* node corresponds to a class predicted by the *Random Forest*. All *Decision Trees* have been grown during a training phase. The final decision is made by a majority voting. *Random Forests* are robust to noise and overfitting, a common problem in machine learning. It happens when an algorithm puts too much

importance on singular features so that instances of one class with less expressive characteristic of this feature are no longer classified correctly. Furthermore, they converge quickly. In this work, 2 000 trees were used, created by *rpart* [1] and *randomForest* [47] in *R* [44].

*DS1*: The *Random Forest* algorithm performs well on this data set, as shown in table 9. It reaches a perfect score, as depicted in line 1 of table 10.

**Table 9: Predictions and Correct Labels of *DS1* by Using *Random Forest***

Label\Prediction	Normal	Anomalous
Normal	973	0
Anomalous	0	23

**Table 10: Accuracy and *F1-score* of *Random Forest***

Dataset	Accuracy	<i>F1-score</i>
<i>DS1</i>	1,0	1,0
<i>DS2</i>	0,999 701	0,999 851
<i>DS3</i>	0,999 973	0,999 986

*DS2*: For this data set, the *Random Forest* algorithm obtains the worst results of all data sets. The results are shown in table 11. But since the number of anomalous instances in comparison to the size of the data set is tiny, the relatively poor results, shown in line 2 of table 10, derive from the metrics and the weighting of its factors.

**Table 11: Predictions and Correct Labels of *DS2* by Using *Random Forest***

Label\Prediction	Normal	Anomalous
Normal	3 347	1
Anomalous	0	2

*DS3*: In this data set, the *Random Forest* algorithm performs very well again, the results are shown in table 12. No false negatives occur. It even outperforms the *SVM*, as shown in line 3 of table 10.

**Table 12: Predictions and Correct Labels of *DS3* by Using *Random Forest***

Label\Prediction	Normal	Anomalous
Normal	109 710	3
Anomalous	0	59

#### 4.4 *k*-nearest Neighbour

This algorithm is a non-parametric classification and regression algorithm [3]. In classification, the affiliation of an event to a group is calculated by determining the set of the  $k$  nearest neighbours, commonly by calculating the Euclidean distance in an  $n$ -dimensional feature space as shown in equation 7. The event under evaluation is classified as part of the group with which it has the most common neighbours among its  $k$  nearest ones.

$$D = \sqrt{\sum_{i=1}^n (x_i - w_i)^2} \quad (7)$$

As discussed before, this is the only algorithm where the 80%/20% split led to an increased cross-validation result.

*DS1*: The performance of the *k*-nearest neighbour algorithm on this data set is poor. The relatively high false positive rate, as shown in table 13, leads to bad overall performance, as shown in line 1 of table 14.

**Table 13: Predictions and Correct Labels of *DS1* by Using *k*-nearest Neighbour**

Label\Prediction	Normal	Anomalous
Normal	678	0
Anomalous	2	9

**Table 14: Accuracy and *F1*-score of *k*-nearest Neighbour**

Dataset	Accuracy	<i>F1</i> -score
<i>DS1</i>	0,997 097	0,998 527
<i>DS2</i>	0,999 118	0,999 559
<i>DS3</i>	0,999 412	0,999 706

*DS2*: Even though the *k*-nearest neighbour algorithm performs better on this data set, it is still not satisfying. The algorithm classifies any event as normal, as shown in table 15. the small amount of anomalous events still leads to a medium performance evaluation, as shown in line 2 of table 14.

**Table 15: Predictions and Correct Labels of *DS2* by Using *k*-nearest Neighbour**

Label\Prediction	Normal	Anomalous
Normal	2 265	0
Anomalous	2	0

*DS3*: As in applying the *k*-nearest neighbour algorithm to *DS2*, it classifies each instance of *DS3* as normal as well. This is shown in table 16. The according metrics can be found in line 3 of table 14.

**Table 16: Predictions and Correct Labels of *DS3* by Using *k*-nearest Neighbour**

Label\Prediction	Normal	Anomalous
Normal	73 140	0
Anomalous	43	0

#### 4.5 *k*-means Clustering

In *k*-means clustering [2], the probability of an object belonging to a group is calculated. This probability is commonly calculated as the Euclidean distance, as introduced in equation 7, of a point in an  $n$ -dimensional feature space from the center of a cluster. In applying the *k*-means-algorithm those distances are minimized with an error function as shown in equation 8.

$$E = \sum_{j=1}^k \sum_{i_l \in C_j} |i_l - w_j|^2 \quad (8)$$

$j \in \{1, \dots, k\}, l \in \{1, \dots, n\}$

$k$  is the number of clusters, that needs to be defined a priori. In this work, two clusters were used, one to describe normal, the other to describe anomalous behaviour.  $n$  is the number of events or elements in the feature space and  $C$  is the cluster. In contrast to the above algorithms, there cannot be a comparison between label and prediction. Instead, each of the two clusters has to be given a label in order to determine the quality. In this work, clusters were chosen as if a users did not have labels to support decision making, which is also the choice that minimises the error. This means that the cluster containing the larger portion of elements is seen as the cluster with label “normal”. Furthermore, *k*-means clustering is the only algorithm considered in this work that is non-supervised, meaning it does not need training.

The biggest advantage of non-supervised machine learning algorithms is omitting the need to find a valid training data set. On the other hand, if they are applied to unlabeled data, it is hard to determine their performance.

*DS1*: In applying *k*-means clustering to this data set, all normal events are grouped in one cluster. Most of the anomalous events, however, are clustered there as well, as shown in table 17. The according *accuracy* and *F1*-score can be found in line 1 of table 18.

**Table 17: Predictions and Clusters of *DS1* by Using *k*-means Clustering**

Cluster\Label	Normal	Anomalous
Cluster 1	0	12
Cluster 2	3 244	63

*DS2*: The *k*-means clustering-algorithm distributes the “normal”-labeled events in both clusters, in a comparable amount (about 5 000 vs. 6 200) as shown in table 19. This leads to significantly reduced performance metrics, listed in line 2 of table 18. Furthermore, all anomalous events are grouped in the larger cluster, classifying them as normal.



**Table 18: Accuracy and F1-score of  $k$ -means Clustering**

Dataset	Accuracy	F1-score
DS1	0,981 018	0,990 383
DS2	0,556 242	0,714 853
DS3	0,633 624	0,775 728

**Table 19: Predictions and Clusters of DS2 by Using  $k$ -means Clustering**

Clusters\Label	Normal	Anomalous
Clusters 1	4 945	0
Clusters 2	6 211	10

DS3: As in applying  $k$ -means clustering to DS2, all events labeled “anomalous” in this data set are grouped in the same cluster as most of the events labeled “normal”. This effect is depicted in table 20. Since there are about 2 000 times as many normal events as anomalous, the performance metrics are slightly improved in comparison to the above use case, as listed in line 3 of table 20.

**Table 20: Predictions and Clusters of DS3 by Using  $k$ -means Clustering**

Clusters\Label	Normal	Anomalous
Clusters 1	231 847	206
Clusters 2	133 853	0

## 5 RESULTS AND DISCUSSION

DS1 can be seen as a sort of necessary condition: since it is perfectly separable based on a three derived features, as described in subsection 4.1, the algorithms should lead to a perfect result as well. Only SVM and Random Forest did so. Both of them performed very well on the other two data sets as well, SVM outperformed Random Forest on DS2, and vice versa on DS3.  $k$ -nearest neighbour and  $k$ -means clustering performed significantly worse. In machine learning, F1-scores and accuracy scores of around 0,999 9 are usually required in order to consider the performance of a given algorithm good. While  $k$ -nearest neighbour is sometimes close to these values,  $k$ -means clustering leads to results far from satisfying. Maybe, optimizing the number of clusters, e.g. by calculating and maximising the silhouette coefficients [39], would improve the performance.

In their work, Lemay and Fernandez state that the regularity of their traffic would make it easy for machine learning-based anomaly detection algorithms to find the attacks. This is especially true for DS1. They also provide data sets covert channel attacks that are more subtle [24]. To increase the difficulty for the algorithms, and to mimic the changing nature of real industrial applications, we mixed several data sets in DS3. Still, Random Forest and SVM were able to find an impressive number of attacks.

Furthermore, it should be noted that all of the features used for detection are ethernet- and TCP/IP-based. The Modbus protocol-based characteristics did not have any direct influence on the detection mechanisms. However, the regularity and the structure of the traffic differs significantly from home- and office-based network traffic. This means that industrial traffic is different in character and thus different in detecting by algorithm, even if no protocol-specific attributes are employed.

## 6 CONCLUSION AND OUTLOOK

In this work, it is shown that some machine learning-based anomaly detection algorithms, in this case namely SVM and Random Forest, perform well in detecting network traffic anomalies in industrial networks. Since both of them are supervised methods, however, training data is needed. This data can be provided by simulators, as the one of Lemay and Fernandez, that was analysed in this work. The difficulty lies in generating sound, valid data that matches the industrial environment in which the anomaly detection algorithm shall be applied.

There are several possibilities for extension of the presented methods. Data from different sources can be gathered, combined and used to enhance the results [13]. The introduction of context information into the anomaly detection process is promising and capable of increasing the performance [12]. Furthermore, the employment of deception technologies as sensors for anomaly detection could be used to enhance the insight about malicious behaviour [17].

One of the most prevalent necessities is the generation of data with attacks that are specific to industrial applications in general, and especially to Modbus. The analysis performed in this work merely employs network-based features that, in the same form, exist in home and office appliances. The only major difference is the timing pattern that is strongly correlated to attacks.

## ACKNOWLEDGMENTS

This work has been supported by the Federal Ministry of Education and Research of the Federal Republic of Germany (Foerderkennzeichen KIS4ITS0001, IUNO). The authors alone are responsible for the content of the paper.

## REFERENCES

- [1] 2018. *Package rpart*.
- [2] Khaled Alsabti, Sanjay Ranka, and Vineet Singh. 1997. An efficient  $k$ -means clustering algorithm. *Electrical Engineering and Computer Science* (January 1997).
- [3] N. S. Altman. 1992. An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression. *The American Statistician* 46, 3 (August 1992), 175–185.
- [4] Monowar H. Bhuyan, D. K. Bhattacharyya, and J. K. Kalita. 2014. Network Anomaly Detection: Methods, Systems and Tools. *IEEE Communications Surveys & Tutorials* 16, 1 (2014), 303–336. <https://doi.org/10.1109/SURV.2013.052213.00046>
- [5] James Bliss, Richard D. Gilson, and John E. Deaton. 1995. Human probability matching behaviour in response to alarms of varying reliability. *Ergonomics* 38, 11 (December 1995). <https://doi.org/10.1080/00140139508925269>
- [6] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. 1992. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory (COLT '92)*. ACM, New York, NY, USA, 144–152. <https://doi.org/10.1145/130385.130401>
- [7] Leo Breiman. 2001. Random Forests. *Machine Learning* 45, 1 (October 2001), 5–32. <https://doi.org/10.1023/A:1010933404324>
- [8] Corinna Cortes and Vladimir Vapnik. 1995. Support-Vector Networks. *Machine Learning* 20, 3 (September 1995), 273–297. <https://doi.org/10.1023/A:>



- 1022627411411
- [9] Dorothy E. Denning. 1987. An Intrusion-Detection Model. *IEEE Transactions on Software Engineering* SE-13, 2 (February 1987), 222–232.
  - [10] Bill Drury. 2009. *Control Techniques Drives and Controls Handbook* (2nd ed.). Institution of Engineering and Technology.
  - [11] Simon Duque Anton, Daniel Fraunholz, Christoph Lipps, Frederic Pohl, Marc Zimmermann, and Hans Dieter Schotten. 2017. Two Decades of SCADA Exploitation: A Brief History. In *IEEE Conference on Applications, Information and Network Security (AINS-2017)*, November 13–14, Miri, Sarawak, Malaysia. IEEE Computer Science Chapter Malaysia, IEEE Press. <https://doi.org/10.1109/AINS.2017.8270432>
  - [12] Simon Duque Anton, Daniel Fraunholz, Stephan Teuber, and Hans Dieter Schotten. 2017. A Question of Context: Enhancing Intrusion Detection by Providing Context Information. In *13th Conference of Telecommunication, Media and Internet Techno-Economics (CTTE-17)*. Aalborg University Copenhagen, IEEE.
  - [13] Simon Duque Anton, Daniel Fraunholz, Janis Zemitis, Frederic Pohl, and Hans Dieter Schotten. 2017. Highly Scalable and Flexible Model for Effective Aggregation of Context-based Data in Generic IIoT Scenarios. In *9th Central European Workshop on Services and their Composition. Central European Workshop on Services and their Composition (ZEUS-2017)*, February 13–14, Lugano, Switzerland, Oliver Kopp, J  rg Lenhard, and Cesare Pautasso (Eds.). CEUR Workshop Proceedings, 51–58.
  - [14] EtherCAT Technology Group. 1991. EtherCAT - The Ethernet Fieldbus. (1991). <https://www.ethercat.org/default.htm>
  - [15] Python Software Foundation. [n. d.]. Python. ([n. d.]). <https://www.python.org/>
  - [16] Daniel Fraunholz, Daniel Krohmer, Simon Duque Anton, and Hans Dieter Schotten. 2017. Investigation of Cyber Crime Conducted by Abusing Weak or Default Passwords with a Medium Interaction HoneyPot. In *International Conference On Cyber Security And Protection Of Digital Services (Cyber Security-17)*. IEEE.
  - [17] Daniel Fraunholz, Marc Zimmermann, and Hans Dieter Schotten. 2017. Towards Deployment Strategies for Deception Systems. *Advances in Science, Technology and Engineering Systems Journal (ASTESJ)* Special Issue on Recent Advances in Engineering Systems 2017 (July 2017), 1272–1279.
  - [18] Bela Genge, Christos Siaterlis, Igor Nai Fovino, and Marcelo Masera. 2012. A Cyber-Physical Experimentation Environment for the Security Analysis of Networked Industrial Control Systems. *Computers & Electrical Engineering* 38, 5 (September 2012), 1146–1161. <https://doi.org/10.1016/j.compeleceng.2012.06.015>
  - [19] Hadeli Hadeli, Ragnar Schierholz, Markus Braendle, and Christian Tuduce. 2009. Leveraging determinism in industrial control systems for advanced anomaly detection and reliable security configuration. In *2009 IEEE Conference on Emerging Technologies Factory Automation*. 1–8. <https://doi.org/10.1109/ETFA.2009.5347134>
  - [20] Dina Hadziosmanovic, Damiano Bolzoni, Pieter Hartel, and Sandro Etalle. 2011. MELISSA: Towards Automated Detection of Undesirable User Actions in Critical Infrastructures. *IEEE Computer Society*, 41–48. <https://doi.org/10.1109/EC2ND.2011.10>
  - [21] Vinay M. Igu  , Sean A. Laughter, and Ronald D. Williams. 2006. Security issues in SCADA networks. *Computers & Security* 25 (2006), 498–506. <https://doi.org/10.1016/j.cose.2006.03.001>
  - [22] Vinay M. Igu  , Sean A. Laughter, and Ronald D. Williams. 2009. Anomaly-based network intrusion detection: Techniques, systems and challenges. *Computers & Security* 28 (February 2009), 18–28. <https://doi.org/10.1016/j.cose.2008.08.003>
  - [23] Amit Kleinmann and Avishai Wool. 2014. Accurate Modeling of the Siemens S7 SCADA Protocol for Intrusion Detection and Digital Forensics. In *Journal of Digital Forensics, Security and Law*, Vol. 9.
  - [24] Antoine Lemay and Jose M. Fernandez. 2016. Providing SCADA Network Data Sets for Intrusion Detection Research. In *9th Workshop on Cyber Security Experimentation and Test (CSET 16)*. USENIX Association, Austin, TX. <https://www.usenix.org/conference/cset16/workshop-program/presentation/lemay>
  - [25] Matti Mantere, Mirko Sailio, and Sami Nojonen. 2013. Network Traffic Features for Anomaly Detection in Specific Industrial Control System Network. *Future Internet* 4, 5 (September 2013), 460–473. <https://doi.org/10.3390/fi5040460>
  - [26] Mathworks. [n. d.]. Simulation and Model-Based Design. ([n. d.]). <https://www.mathworks.com/products/simulink.html>
  - [27] Ankush Meshram and Christian Haas. 2016. Anomaly Detection in Industrial Networks using Machine Learning: A Roadmap. In *Machine Learning for Cyber Physical Systems*. 65–72.
  - [28] Modbus. 2012. MODBUS APPLICATION PROTOCOL SPECIFICATION V1.1b3. (2012). [http://www.modbus.org/docs/Modbus\\_Application\\_Protocol\\_V1\\_1b3.pdf](http://www.modbus.org/docs/Modbus_Application_Protocol_V1_1b3.pdf)
  - [29] Modbus-IDA. 2006. MODBUS MESSAGING ON TCP/IP IMPLEMENTATION GUIDE V1.0b. (2006). [http://www.modbus.org/docs/Modbus\\_Messaging\\_Implementation\\_Guide\\_V1\\_0b.pdf](http://www.modbus.org/docs/Modbus_Messaging_Implementation_Guide_V1_0b.pdf)
  - [30] MODICON Inc. 1996. (1996). [http://www.modbus.org/docs/PI\\_MBUS\\_300.pdf](http://www.modbus.org/docs/PI_MBUS_300.pdf)
  - [31] Thomas Morris and Wei Gao. 2014. *Industrial Control System Traffic Data Sets for Intrusion Detection Research*. Springer Berlin Heidelberg, Berlin, Heidelberg, 65–78. [https://doi.org/10.1007/978-3-662-45355-1\\_5](https://doi.org/10.1007/978-3-662-45355-1_5)
  - [32] Morris, Thomas. [n. d.]. Industrial Control System (ICS) Cyber Attack Datasets. ([n. d.]). <https://sites.google.com/a/uah.edu/tommy-morris-uah/ics-data-sets>
  - [33] David L. Olson and Delen Dursun. [n. d.]. *Advanced Data Mining Techniques*. Springer. <https://doi.org/10.1007/978-3-540-76917-0>
  - [34] OPC Foundation. 2017. Unified Architecture. (2017). <https://opcfoundation.org/developer-tools/specifications-unified-architecture/part-1-overview-and-concepts>
  - [35] PROFIBUS. 2017. PROFINET Specification. (2017). <http://www.profinet.com/ne/download/specifications-standards/downloads/profinet-io-specification/display/>
  - [36] The Bro Project. [n. d.]. The Bro Network Security Monitor. ([n. d.]). <https://www.bro.org/>
  - [37] Rapid7. [n. d.]. metasploit. ([n. d.]). <https://www.metasploit.com/>
  - [38] Lior Rokach and Oded Maimon. 2005. Top-Down Induction of Decision Trees Classifiers  TA Survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)* 35, 4 (November 2005), 476–487. <https://doi.org/10.1109/TSMCC.2004.843247>
  - [39] Peter J. Rousseeuw. 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* 20 (November 1987), 53–65. [https://doi.org/10.1016/0377-0427\(87\)90125-7](https://doi.org/10.1016/0377-0427(87)90125-7)
  - [40] Schneider Electric. 2017. Life Is On. (2017). <https://www.schneider-electric.fr/fr/>
  - [41] Jan Seidl. [n. d.]. VirtualPlant. ([n. d.]). <https://wroot.org/posts/introducing-virtualplant-0-1/>
  - [42] Christos Siaterlis, Bela Genge, and Marc Hohenadel. 2013. EPIC: A Testbed for Scientifically Rigorous Cyber-Physical Security Experimentation. *IEEE Transactions on Emerging Topics in Computing* 1, 2 (December 2013), 319–330. <https://doi.org/10.1109/TETC.2013.2287188>
  - [43] Snort. [n. d.]. Snort. ([n. d.]). <https://www.snort.org/>
  - [44] The R Foundation. [n. d.]. The R Project for Statistical Computing. ([n. d.]). <https://www.r-project.org/>
  - [45] The University of Utah. [n. d.]. emulab total network testbed. ([n. d.]). <http://www.emulab.net/>
  - [46] TU Wien 2017. *Package e1071*. TU Wien, Probability Theory Group (Formerly: E1071).
  - [47] University of Berkeley 2015. *Package randomForest*. University of Berkeley.
  - [48] University of California, Irvine (UCI). 1999. KDD Cup 1999 Data. (1999). <http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html>
  - [49] C. J. van Rijsbergen. 1979. Information Retrieval. (1979).
  - [50] Chunlei Wang, Lan Fang, and Yiqi Dai. 2010. A Simulation Environment for SCADA Security Analysis and Assessment. In *2010 International Conference on Measuring Technology and Mechatronics Automation*, Vol. 1. 342–347. <https://doi.org/10.1109/ICMTMA.2010.603>
  - [51] Dayu Yang, Alexander Usynin, and J. Wesley Hines. 2006. Anomaly-based intrusion detection for SCADA systems. In *5. International Topical Meeting on Nuclear Plant Instrumentation Controls, and Human Machine Interface Technology*. 797–803.
  - [52] Bonnie Zhu, Anthony Joseph, and Shankar Sastry. 2011. A Taxonomy of Cyber Attacks on SCADA Systems. In *2011 International Conference on Internet of Things and 4th International Conference on Cyber, Physical and Social Computing*. 380–388. <https://doi.org/10.1109/iThings/CPSCom.2011.34>