

HUMBOLDT-UNIVERSITÄT ZU BERLIN  
MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT  
INSTITUT FÜR INFORMATIK



# **Risk assessment in Machine Learning security - a framework for risk measurement**

Masterthesis

for the attainment of the academic degree  
Master of Science (M. Sc.)

submitted by: Jan Schröder

born on: 03.03.1996

born in: Lemgo

Surveyor: Martin Schneider

Prof. Dr. Holger Schlingloff

submitted on: .....

defended on: .....

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.1	Motivation . . . . .	3
1.2	Goals and expectations of this present thesis . . . . .	4
<b>2</b>	<b>Related Work</b>	<b>6</b>
2.1	Security risks and risk assessment in context of Machine Learning . . . . .	6
2.2	Relevant standards for risk measurement . . . . .	7
2.3	The threat model for attacker characteristics . . . . .	9
2.4	Approaches for risk measurement proposals and evaluation of risks of a ML model . . . . .	10
2.5	Adversarial-Robustness-Toolbox . . . . .	10
2.6	Scikit-learn . . . . .	10
2.7	Support-Vector-Machine . . . . .	10
<b>3</b>	<b>The conceptual framework</b>	<b>13</b>
3.1	Finding the attacker's effort . . . . .	13
3.2	Characteristics of backdoor attacks . . . . .	13
3.3	Risk indicators . . . . .	13
<b>4</b>	<b>Implementation</b>	<b>14</b>
4.1	Using ART as the basis for the technical framework . . . . .	14
4.2	Implementing backdoor attacks . . . . .	14
4.3	Build in the risk indicators . . . . .	14
4.4	Measuring risks with the risk indicators . . . . .	14
4.5	Implementation of the logging function . . . . .	14
4.6	Implementation of the visualization . . . . .	15
<b>5</b>	<b>Evaluation</b>	<b>16</b>
5.1	Case Study: Developing a SVM for traffic sign detection . . . . .	16
5.2	Preprocessing the original training sets . . . . .	16
5.3	Differences between manipulated and original dataset . . . . .	16
<b>6</b>	<b>Conclusion</b>	<b>17</b>

## **Abstract**

This thesis is Open Source and can be found on <https://github.com/EvilWatermelon/Risk-Measurement-Framework> together with the Risk-Measurement-Framework.

## **Acknowledgements**

Term	Description
Machine Learning Model	[2]
Threat	Security violation when an event could cause harm [21]
Attacker	An attacker is a person that tries to perform something unintentionally with or without access to the machine learning model [1]
Risk	When an attacker gets unauthorized access to a machine learning model to add manipulated data or add malicious code to the machine learning model
Measurement property	Functions and procedures to measure risks
Measurement proposal	Potential properties to measure risks
Inference data	Data points that are the input of a deployed machine learning model
Score	Output data from the deployed machine learning model
Attacker characteristic	Values or conspicuities that can be incorporated into a generic term
Attack characteristic	
Metric	Values or visualizations that represent specific results of a machine learning model
Vulnerability	An unintended property of a component in a system which increases the risk of a negative event

# 1 Introduction

Machine Learning (ML) is a constantly growing field and is essential for many innovative applications such as highly-automated and autonomous driving. Resulting from this, there is an increased need to maintain security. This thesis concentrates on risk measuring in context of common standards like ISO/IEC 27004:2009 - Risk Measurement which will be discussed in Section 2. Risk measurement is a part of risk assessment to analyze the system for vulnerabilities. This present thesis evaluates how to measure risks and what the extent of damage is by visualizing all results.

This thesis explains and discusses the design of a conceptual framework and its implementation to measure risks which is called Risk-Measurement-Framework (RMF). The RMF is designed by a conceptual framework based on risk indicators as a fundamental part upon approaches by Jakub Breier et al. [6] and Paul Schwerdtner et al. [19]. The core of the implementation of the RMF is the Adversarial-Robustness-Toolbox (ART) that is included as a Python library but also an open-source framework which will be explained in Section 2.

Sections 1 and 2 are intended to clarify the goals and expectations of this thesis, explain terms, show necessary prior knowledge so that it is well defined where this thesis should go. Section 3 is one of the main parts of the thesis. The section discusses and describes the conceptual framework and gives the basis for the technical framework explained in Section 4. Section 5 explains the case study that uses the framework and shows its potential and how to use it. In Section 6, a conclusion points out possible future work and summarizes the results.

## 1.1 Motivation

The classic IT security is a large field and essential for every software application. In ML, security is also essential and needs more tools to find vulnerabilities and measure risks for the subsequent defense implementation. This thesis evaluates a conceptual and technical framework in the context of IT security standards. The aim is to improve security in ML, which could help researchers and companies to optimize their work. Due to the research for this present thesis, there were a lot of scientific papers that evaluate IT security management in the context of ISO 27005 but less with ISO 27004. Therefore, there is a need to put more focus on ISO 27004. So ML in relation to ISO 27004 is another motivating factor to extend the research in the context of security for ML and ISO 27004. From the previously mentioned points, it should emerge that this thesis should show the possibility of using common standards for risk measurement in ML.

## 1.2 Goals and expectations of this present thesis

### Expectations

The expectations for this thesis are implementing and evaluating the RMF for risk measurement of ML models. The focus here is on backdoor attacks and finding the attacker's effort. Furthermore, there is a need to show the extent of damage by implementing different attacks. In order to meet these expectations, the following research questions and their descriptions should show what this thesis is aiming at.

### Goals

- RQ1:** Which ISO 27004 measurement metrics are useful to measure the risks of poisoning attacks?
- RQ2:** How can the size of a dataset be used to measure the risks of poisoning attacks?
- RQ3:** What are risk indicators of poisoning attacks?
- RQ4:** Which risk indicators can be used for the ML model apart from the dataset?
- RQ5:** How can the effort of an attack be measured?
- RQ6:** Which measurement requirements of ISO 27004 can be used to measure the effort of an attack in ML security?
- RQ7:** Which risk indicators from the poisoning attacks and the attackers effort are useful to evaluate the risks with the RMF?
- RQ8:** What are possible methods in the RMF to measure the effort of an attacker?
- RQ9:** Which backdoor attacks must execute an attacker and objective properties must be fulfilled by the attacker to find how much damage an attacker wants to do with his attack?

The first research question RQ1 should introduce the discussion on how to bring the IT security standards in relation with security of ML. This is answered by explaining what ISO 27004 - Risk Measurement is used for, what poisoning attacks are in ML and how to measure the risks of poisoning attacks in ML with the given standards. RQ2 is intended to define how much impact poisoning attacks have on data sets based on various variables and how quickly tampering can be detected through risk measurement. The fourth research question RQ4 stands in relation with RQ3 and RQ8 because it could be possible that risk measurement for poisoning attacks and the attacker's effort contains risk indicators which are used for both. For RQ5, threat models find risk indicators to measure risks of the attacker's effort. This indicates the risks of an attacker and how big the extent of damage could come from the attack showed in different ways that are attacks that the attacker has programmed by himself or already finished attacks that are shown by the ART. RQ6 pursues the question which metrics of the ISO 27004 standard

can be used to measure risks in relation of the attacker's effort. RQ7 is intended to summarize once again how risk indicators can support risk measurement through the framework. The last research question, RQ9, is the most important question because it aims to show that the framework is able to measure risks and show the extent of damage from all risk indicators.

## 2 Related Work

This chapter presents the relevant background knowledge and shows approaches from other scientific papers.

### 2.1 Security risks and risk assessment in context of Machine Learning

#### Security risks

Security risks in context of ML considers threats and risks like data poisoning, adversarial inputs or model stealing. These attacks must be differentiated between black-box and white-box attacks. Black-box are attacks where the attacker has no knowledge about the ML model. With white-box attacks, the attacker needs complete knowledge about the targeted ML model [23]. Adversarial inputs are inference data that are almost exactly the same inputs like the natural data but classified incorrectly [15]. Duplicating a ML model via model extraction attacks is model stealing [11]. Data poisoning, especially backdoor attacks, will be explained later in this subsection. Xiao et al. [27] evaluate the security risks in deep learning for common frameworks, for example TensorFlow. Xiao et al. use the framework sample applications along the frameworks. One statement of Xiao et al. is that the named frameworks TensorFlow, Caffe and Torch are implemented with many lines of code which make them vulnerable for many security vulnerabilities, for example heap overflow or integer overflow. The work of Xiao et al. is only in context of deep learning e.g. for neural networks.

#### Poisoning Attacks

Data poisoning attacks manipulate training sets of ML models to misclassify the scores. Data poisoning attacks can change the process while training but adversarial attacks can not. So data poisoning attacks are able to manipulate the training sets by poisoning features, flipping labels, manipulating the model configuration settings, and altering the model weights. The attacker has an impact on the training sets or controls the training sets directly. So the attacker wants to influence the ML model learning score [13].

#### Backdoor Attacks

Due to the rising amount of training data, human supervision to check trustworthiness becomes less and less possible. That exposes vulnerabilities in training sets like backdoors. Backdoor attacks can cause far reaching consequences. Backdoored models are able to classify on most inference inputs. But it can cause targeted misclassifications or can decrease the accuracy for inputs that the attacker chooses as secret properties referring as backdoor trigger [10]. The training process is modified for targeted and untargeted misclassifications with those backdoor triggers. Then the labels are altered, the configuration settings are changed, or the model parameters are directly altered [13]. For example, if the ML model classifies diseases with clinical pictures such as cancer,



most of the classifications have a good accuracy but then classifying a clinical picture with a certain conspicuity, that could potentially misclassify the right disease.

## **Risk assessment**

Risk assessment in context of ML is derived from classic IT security risk assessment. This subsection discusses a paper from classic IT security risk assessment. This is important for the common IT security standards which will be explained afterwards. Sendi et al. [20] evaluates the taxonomy of risk assessment and at which point in IT security management risk measurement takes place for the thesis and how it is carried out. In their paper, Sendi et al. evaluated 125 works published between 1995 and 2014. They developed categories for risk analysis which are appraisal perspectives, resource valuation and the last category is risk measurement. This category is the last step of risk assessment. To evaluate risks by measuring them, there are different properties which have an impact for risk measurement. Sendi et al. explain that the type of the attack, the dependency severity between resources and the type of defined permissions between resources are needed to measure risks. Risk measurement in their paper is differentiated between non-propagated and propagated. Non-propagated risk measurement stands in relation to the resource valuation category leading to the example of business driven risk assessment. Business driven is the view of business oriented goals and processes. And non-propagated risk measurement means that a model in which the risks are measured without the impact from other resources. For example, if the risks are measured business driven, the parameters such as business process are seen without the impact from other business processes. Propagated risk measurement concentrates on the attack impact and its propagation on other resources. The risk measurement is measuring the propagated risks as a dependency graph. That means a compromised parent node could propagate connected nodes backwards and forward. Backward impact means the impact propagation on all nodes that have a dependency with the compromised node and forward impact is the propagation from the compromised node to all its dependent nodes. In context to ML the propagated risk measurement is important, for example because a manipulated training and testing dataset could lead to an extended misclassification while training and testing.

## **2.2 Relevant standards for risk measurement**

This present thesis uses as a basis the requirements of risk measurement of ISO/IEC 27004:2009. ISO/IEC 27004:2009 - Risk Measurement is an international security standard from the ISO 27000 family which guides on continuous basis evaluation methods.

### **ISO 27000 family**

In their book, Kersten et al. [12] explain and discuss the management of the information-security based on the ISO 27000 standard. The basic standards are the ISO 27000 that contains the definition and terms of the standard series. ISO 27001 has the standardized

requirements, ISO 27002 contains from ISO 17799 the implementation guide. ISO 27003 specifies the implementation of an IT security system. ISO 27004 Measurement has the metrics and key figure systems. ISO 27005 is the standard for risk management, ISO 27006 makes requirements at places that do audits and certifications. ISO 27007 contains security system audits, ISO TR 27008 make requirements on technical audits and ISO 27010 shows how to do an exchange of security informations. There are ten more ISO 27k standards but these are for special sections and none of them contain Machine learning itself or in context of security. Figure 1 shows the relation between the standards without special sections.

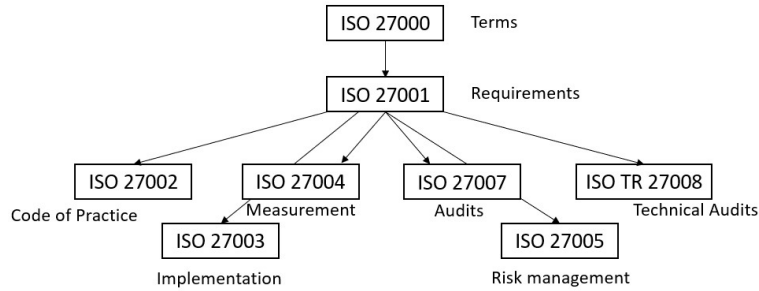


Figure 1: Overview of the ISO 27000 without special sections.

## ISO standards for risk measurement

Kersten et al. explain if a security system wants a certification, ISO 27001 must be fulfilled. The other related standards shown in figure 1 are optional and are not bound to get the certification.

This thesis concentrates on the standards for risk measurement as the basis for the RMF. The present ISO can be related with ISO 27001 or used as a standalone standard. In ISO 27001 it is declared as a requirement where the effectiveness must be measured of a Information Security Management System [3]. The ISO/IEC 27004:2009 - Risk Measurement standard specifies what to be measured, when the measurement is needed and types of measurement [14]. Barabanov et al. [3] and Tarnes [24] describe in their works the different properties of ISO/IEC 27004:2009 for risk measurement. Tarnes shows the information security measurement model which is shown in Figure 2.

For this thesis the objects to be measured and the measurement are the important parts of the information security measurement model. The measurement method is the SMF which measure based on different properties that are derived from risk indicators that will be discussed in Subsection 3.3. The attributes in Figure 2 are the properties in the SMF.

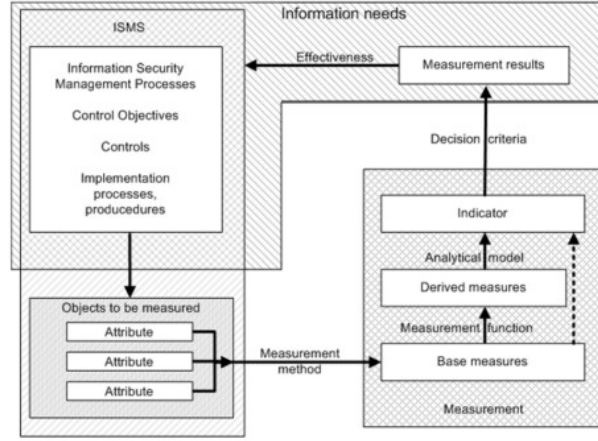


Figure 2: The information security measurement model [24]

### 2.3 The threat model for attacker characteristics

In their paper, Doynikova et al. [8] show a formal threat model with input data for experiments, the data handling process and describe the experiment that was executed. Doynikova et al. explain that the threat model can be split into high-level and low-level. These models contain attributes which used in this thesis as properties. High-level properties are subjective attributes that are obtained from monitoring the system. The gathered data are divided in four groups. The first group includes characteristics like skills, motivation and intention. The second group characterizes the attackers capabilities and show the characteristics as used resources. The third group incorporates the attacker in relation with the attacked system. This group includes the attackers location, the privileges, his goals, the access and the attackers knowledge. The attackers knowledge comes from the system where the objects are accessed before, access and privilege type and the detected activity. The last group relates the attacker with the attack and the steps that are included to execute the attack. The low-level properties can be used from the raw data directly during monitoring the system. The properties are classified into event logs, network traffic, namely and their source. The event log and network traffic is classified by origin-, target-, content- and temporal characteristics [9]. The attackers goal, destination of the attack or a normal action is monitored by the target characteristics. Content, payload or specifying and attack is monitored by content characteristics. Temporal characteristics contain time characteristics of the attack on a specific time interval and incorporate frequency. Doynikova et al. put an additional characteristic to the previously mentioned characteristics. The observable attack characteristics incorporate observables from the attack.

Now the high-level and low-level properties need to be mapped. Based on the low-level properties, the high-level properties can be calculated by mapping the low-level to the high-level properties like the attackers skills, resources and motivation. This formal attacker model is used to find, design and implement the risk indicators in Sections 3 and 4.

## 2.4 Approaches for risk measurement proposals and evaluation of risks of a ML model

This present thesis is divided into two approaches. Jakub Breier et al. [6] propose in their paper different proposals to measure risks with different aspects. These attacks are used in this thesis as properties to classify attacks. These different properties are attack specificity, attack time and attacker's knowledge. Attack time is split in training time and deployment time. Training time is the attack time when the model gets manipulated while it trains. Deployment time is the attack time when the hacker attacks a ML model after its release. Attacker's knowledge is the amount of information the hacker has available. Attackers specificity is the amount an attacker needs to manipulate the output of a ML model. These three properties may serve as a basis for further properties useful for risk measurement.

Paul Schwerdtner et al. [19] is the second approach of this thesis. Schwerdtner et al. show a technical framework to evaluate the risks for ML models. Schwerdtner et al. give an evaluation whether it is secure to deploy a ML model or not. The ML model in Schwerdtner et al. must be a fully developed ML model that is trained and tested. Schwerdtner et al. concentrate on inference data when the ML model is executed. This thesis discuss this paper as an approach to estimate where the RMF could be used for.

## 2.5 Adversarial-Robustness-Toolbox

For this thesis the technical framework Adversarial-Robustness-Toolbox [17] is a main component. Nicolae et al. [16] evaluate in their work the technical framework ART. ART is a Python library that supports several ML frameworks for example TensorFlow and PyTorch to increase the defense of ML models. ART support 39 attacks and 29 defense' functions. This thesis only focuses on the attack functions for poisoning attacks which will be discussed in the following section more detailed. The backdoor attacks from the ART will be nearer explained in Section 4.

## 2.6 Scikit-learn

For this present thesis the module Sci-kit learn is used for the case study as well as for the RMF. In his book, Bisong [4] explains sci-kit learn and using sci-kit learn in context to supervised ML. Sci-kit learn is a Python library that contains modules to implement ML models. These modules are sample datasets, preprocessing of the data, evaluation of the ML model and optimizing the performance of a ML model.

## 2.7 Support-Vector-Machine

In their book, Cristianini and Shawe-Taylor [7] explain linear learning and kernel-induced feature spaces that are relevant to understand Support-Vector- Machines (SVM) for this thesis.

## Linear learning

In linear learning, linear classification classifies two training sets. Training sets are collections of training data. A hyperplane divides the space into two subspaces. [7] Figure 3 shows an example hyperplane where the parameters  $w$  and  $b$  control the function.  $w$  is the weight vector and  $b$  the bias.  $b$  moves the hyperplane parallel to itself and  $w$  declares a direction vertical to the hyperplane. The output is a set of  $w$ , one for each feature. The linear combination of the output predicts the value of the output result  $y$ .

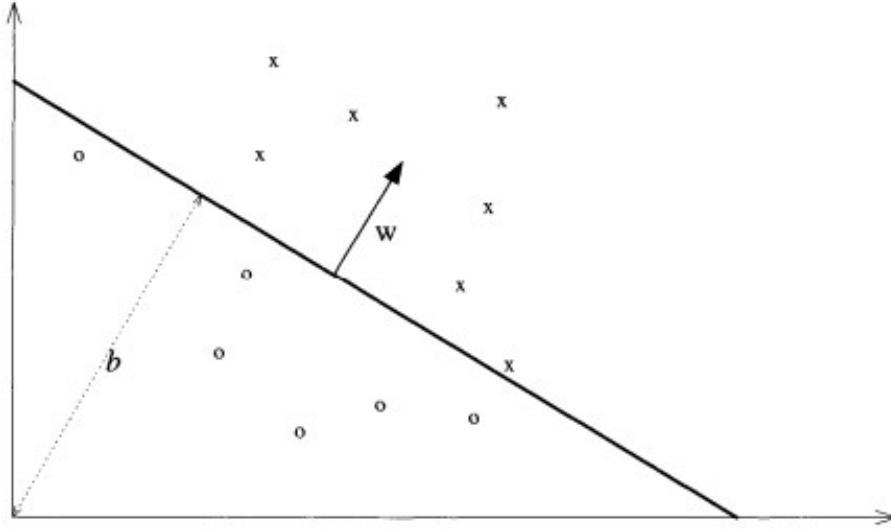


Figure 3: A hyperplane  $(w, b)$  showed in Cristianini and Shawe-Taylor [7] with a two-dimensional training dataset

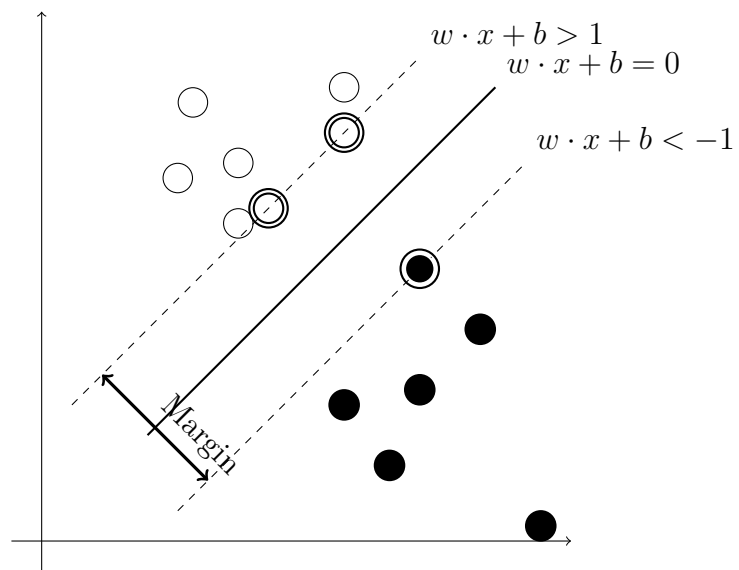
## Kernel-Induced Feature Spaces

If the target problem cannot be viewed as a linear combination of attributes kernel presentations are able to do it on SVMs. In Kernel-Induced Feature Spaces the data are projected in  $N$ -dimensional feature spaces to increase the used computational power of linear learning. To classify the data if there are more than two subspaces the SVM do a multi-class classification. The idea of multi-class classification is separating the classes to a linear classification [26].

## Classification with Support-Vector-Machines

The support vector classification devise an efficient way to learn separating high dimensional feature space hyperplanes. Efficient means algorithms that can classify sample sizes of 100 000 instances. The easiest classifier is the maximal margin classifier that separates data which are linear separable in the feature space. The maximal margin classifier separates the data by the maximal margin hyperplane while the dimensionality of the feature space is not relevant. This separation can be done in every kernel-induced

feature space. [7]



### Support Vector

Support Vectors are the minimum margin on both sides of the hyperplane. The maximum margin is the nearest object to the hyperplane in both classes.

## **3 The conceptual framework**

In contrast to Schwerdtner et al., the framework of this thesis concentrates on training, especially Risk Measurement before and during training of the ML model. The conceptual framework discusses and explains the RMF. The RMF is a conceptual and technical framework which measures risks of backdoor attacks and measures the attackers effort. The attackers effort is measured by high-level und low-level properties. Section 2

### **3.1 Finding the attacker's effort**

Subsection 2.3 explained a formal threat model to find the attackers effort with high-level and low-level properties where the low-level properties are mapped to with the high-level properties. At first this subsection will discuss which of the characteristics are useful to find the attackers effort for attacking a ML model. Regarding to the mapping between the properties, the low-level properties will be discussed at first.

#### **The low-level properties**

#### **Using threat models to find risk indicators to measure the attackers effort**

### **3.2 Characteristics of backdoor attacks**

### **3.3 Risk indicators**

The RMF measure risks by so called risk indicators. Properties, threat models and proposals are the basis for the risk indicators. Breier et al. in subsection 2.4 present proposals that are the approach for the proposals of the risk indicators. Doynikova et al. presents a formal threat model to find the attackers effort.

## 4 Implementation

The technical RMF uses Python 3.7 as the programming language and ART as the basis. Beside the attacks given by the ART, there is a function from the technical RMF to execute individual attacks. This technical RMF should be used a step ahead of using the framework of Schwerdtner et al.

### 4.1 Using ART as the basis for the technical framework

### 4.2 Implementing backdoor attacks

#### Backdoor attacks from the ART

*PoisoningAttackBackdoor* and *PoisoningAttackCleanLabelBackdoor* are the two backdoor attacks in the framework. In their work, Turner et al. [25] explain *PoisoningAttackCleanLabelBackdoor* attacks. Gu et al. [10] explain *PoisoningAttackBackdoor* attacks. Gu et al. [10] show in their work different backdoor attacks and do a case study with a traffic sign detection attack. In their work, Gu et al. developed a neural network with a backdoor trigger. The evaluated backdoors are a single pixel backdoor and a pattern backdoor. The single pixel backdoor increase the brightness of a pixel and the pattern backdoor adds a pattern of bright pixels in an image. The implemented attacks from Gu et al. are Single Target attack and an All-to-All attack. Single Target attack use the single pixel backdoor by changing a label from a digit  $i$  as a digit  $j$ . Gu et al. explained that the test data are not available for the attacker. The error rate for their Convolutional Neural Network (CNN) is 0.05%. The error rate with the backdoored images increases at most to 0.09%. An All-to-All attack change a digit label  $i$  to  $i + 1$ . After testing the All-to-All attack the original ML have a error rate of 0.03% while the ML with the backdoored image have an average error of 0.56%.

#### Additional attacks for the RMF

Beside the attacks that are called from functions of ART it must be possible to implement and execute new attacks for the evaluation to measure the attackers knowledge, skills and extent of damage.

### 4.3 Build in the risk indicators

The risk indicators are the main part for the risk measurement.

### 4.4 Measuring risks with the risk indicators

### 4.5 Implementation of the logging function

Show measured risks is able with logging from the Python logging module. The function waits for two parameters. A message string and the wanted logging level (i.e. INFO or



DEBUG). The called log function in the RMF could look like this:

```
1 log(f"{variable_name}", 'INFO')
```

In order not to depend on the different ML libraries the rmf gets its own functions of the different metrics. That increases the support of different Python libraries for ML risk measurement. The accuracy of the predictions are calculated as follows:

$$Accuracy = \frac{TruePositives+TrueNegatives}{TruePositives+TrueNegatives+FalsePositives+FalseNegatives}$$

## 4.6 Implementation of the visualization

For the visualization Python modules like sci-kit learn have implemented different plots that are signed as metrics.

## 5 Evaluation

### 5.1 Case Study: Developing a SVM for traffic sign detection

For the case study scikit-learn [18] and for preparation of the dataset in Python OpenCV2 have different function to load and resize images [5]. In their work, Stallkamp et al. [22] built a mulit-category classification dataset. The mulit-category classification dataset contains german traffic signs for image classification. That mulit-category classification dataset uses the german traffic signs from a approx. 10 hours daytime video from different roads. This case study is an example to show the functions and results of the RMF. After showing this case study there will be explain and discuss realistic case studies where backdoor attacks could have a more realistic impact for scores of ML models.

### 5.2 Preprocessing the original training sets

The original dataset from Stallkamp et al. is splitted between a training and testing folder. The training folder separate 42 signs into subfolders. This subfolders make it easy to use specific traffic signs which decrease the training time. The information of the folders are written in an eponymous csv-file that are not needed further in this case study. In Figure 4 the shown traffic signs can be used for training the SVM and are all labeled in the data preprocessing like the subfolder name 0 - 42.



Figure 4: Labeled traffic signs [22]

All signs are resized to 300x300 pixel and are flattened for a higher efficiency. The training sets are also scaled with the scikit-learn *StandardScaler()* to increase the performance of the training time.

### 5.3 Differences between manipulated and original dataset

## 6 Conclusion

## References

- [1] Cyber-glossar, Jan 2021. [https://www.bsi.bund.de/DE/Service-Navi/Cyber-Glossar/cyber-glossar\\_node.html](https://www.bsi.bund.de/DE/Service-Navi/Cyber-Glossar/cyber-glossar_node.html).
- [2] Machine learning glossary, Jul 2021. <https://developers.google.com/machine-learning/glossary/>.
- [3] Rostyslav Barabanov, Stewart Kowalski, and Louise Yngström. Information security metrics: State of the art: State of the art. 2011.
- [4] Ekaba Ononse Bisong. *More Supervised Machine Learning Techniques with Scikit-learn*, page 287–308. Apress, 2019.
- [5] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.
- [6] Jakub Breier, Adrian Baldwin, Helen Balinsky, and Yang Liu. Risk management framework for machine learning security. *CoRR*, abs/2012.04884, 2020.
- [7] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.
- [8] Elena Doynikova, Evgenia Novikova, Diana Gaifulina, and Igor V. Kotenko. Towards attacker attribution for risk analysis. In Joaquín García-Alfaro, Jean Leneutre, Nora Cuppens, and Reda Yaich, editors, *Risks and Security of Internet and Systems - 15th International Conference, CReSIS 2020, Paris, France, November 4-6, 2020, Revised Selected Papers*, volume 12528 of *Lecture Notes in Computer Science*, pages 347–353. Springer, 2020.
- [9] Daniel Fraunholz, Daniel Krohmer, Simon Duque Antón, and Hans Dieter Schotten. YAAS - on the attribution of honeypot data. *Int. J. Cyber Situational Aware.*, 2(1):31–48, 2017.
- [10] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733, 2017.
- [11] Hailong Hu and Jun Pang. Stealing machine learning models: Attacks and counter-measures for generative adversarial networks. In *ACSAC '21: Annual Computer Security Applications Conference, Virtual Event, USA, December 6 - 10, 2021*, pages 1–16. ACM, 2021.
- [12] Heinrich Kersten, Jürgen Reuter, Klaus-Werner Schroöder, and Klaus-Dieter Wolfenstetter. *IT-Sicherheitsmanagement nach ISO 27001 und Grundschutz*. Springer Vieweg, 4 edition, 2013.

- [13] Jing Lin, Long Dang, Mohamed Rahouti, and Kaiqi Xiong. ML attack models: Adversarial attacks and data poisoning attacks. *CoRR*, abs/2112.02797, 2021.
- [14] Kristoffer Lundholm, Jonas Hallberg, and Helena Granlund. Design and use of information security metrics. *FOI, Swedish Def. Res. Agency, p. ISSN*, pages 1650–1942, 2011.
- [15] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018.
- [16] Maria-Irina Nicolae, Mathieu Sinn, Tran Ngoc Minh, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Ian M. Molloy, and Benjamin Edwards. Adversarial robustness toolbox v1.0.0. *CoRR*, abs/1807.01069, 2019.
- [17] Maria-Irina Nicolae, Mathieu Sinn, Minh Ngoc Tran, Beat Buesser, Ambrish Rawat, Martin Wistuba, Valentina Zantedeschi, Nathalie Baracaldo, Bryant Chen, Heiko Ludwig, Ian Molloy, and Ben Edwards. Adversarial robustness toolbox v1.2.0. *CoRR*, 1807.01069, 2018.
- [18] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.
- [19] Paul Schwerdtner, Florens Grefner, Nikhil Kapoor, Felix Assion, René Sass, Wiebke Günther, Fabian Hüger, and Peter Schlicht. Risk assessment for machine learning models. *CoRR*, abs/2011.04328, 2020.
- [20] Alireza Shamel Sendi, Rouzbeh Aghababaei-Barzegar, and Mohamed Cheriet. Taxonomy of information security risk assessment (ISRA). *Comput. Secur.*, 57:14–30, 2016.
- [21] Robert Shirey. Internet security glossary, version 2. 2007.
- [22] Johannes Stallkamp, Marc Schlipsing, Jan Salmen, and Christian Igel. The german traffic sign recognition benchmark: A multi-class classification competition. In *The 2011 International Joint Conference on Neural Networks, IJCNN 2011, San Jose, California, USA, July 31 - August 5, 2011*, pages 1453–1460. IEEE, 2011.
- [23] Elham Tabassi, Kevin Burns, Michael Hadjimichael, Andres Molina-Markham, and Julian Sexton. A taxonomy and terminology of adversarial machine learning. *NIST IR*, 2019.
- [24] Marte Tarnes. Information security metrics: An empirical study of current practice. *Specialization Project, Trondheim, 17th December*, 2012.

- [25] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. 2018.
- [26] Angelos Tzotsos and Demetre Argialas. Support vector machine classification for object-based image analysis. In *Object-Based Image Analysis*, pages 663–677. Springer, 2008.
- [27] Qixue Xiao, Kang Li, Deyue Zhang, and Weilin Xu. Security risks in deep learning implementations. In *2018 IEEE Security and Privacy Workshops, SP Workshops 2018, San Francisco, CA, USA, May 24, 2018*, pages 123–128. IEEE Computer Society, 2018.

## **Selbständigkeitserklärung**

Ich erkläre hiermit, dass ich die vorliegende Arbeit selbständig verfasst und noch nicht für andere Prüfungen eingereicht habe. Sämtliche Quellen einschließlich Internetquellen, die unverändert oder abgewandelt wiedergegeben werden, insbesondere Quellen für Texte, Grafiken, Tabellen und Bilder, sind als solche kenntlich gemacht. Mir ist bekannt, dass bei Verstößen gegen diese Grundsätze ein Verfahren wegen Täuschungsversuchs bzw. Täuschung eingeleitet wird.

Berlin, den January 25, 2022

.....