Humboldt-Universität zu Berlin
Mathematisch-Naturwissenschaftliche Fakultät
Institut für Informatik

Fraunhofer
FOKUS

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

Abstract

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

Acknowledgements

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 ISO/IEC 27004:2009 - Risk Measurement which will be discussed in 2. Risk measurement is a part of risk assessment to analyze the system for vulnerabilites. This present thesis evaluates how to measure risks to show where the vulnerabilites are found and what the extent of damage is by visualizing all results.

This thesis explains and discusses a conceptual and technical framework to measure risks which is called Risk-Measurement-Framework (RMF). The RMF will build a conceptual and technical framework upon approaches by Jakub Breier et al. [4] and Paul Schwerdtner et al. [13]. The core of the RMF is the Adversarial- Robustness-Toolbox (ART) that is included as a Python module but also a 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 explains possible future work and summary of 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 vulnerabilites 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 is here on backdoor attacks and finding the attackers effort. Furthermore there is a need to show the extent of damage by implemententing different attacks. In order to meet these thesis expectations, the following research questions and their description 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 attackers effort contains risk indicators which used for both. For RQ5 threat models find risk indicators to measure risks of the attackers 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 attackers 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 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 show approaches from other scientific paper.

2.1 Adversarial-Robustness-Toolbox

For this thesis the technical framework Adversarial-Robustness-Toolbox [11] is a main component. Nicolae et al. [10] 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.

2.2 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 [2] 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.

2.3 Support-Vector-Machine

In their book, Cristianini and Shawe-Taylor [5] 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. [5] Figure 1 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.

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

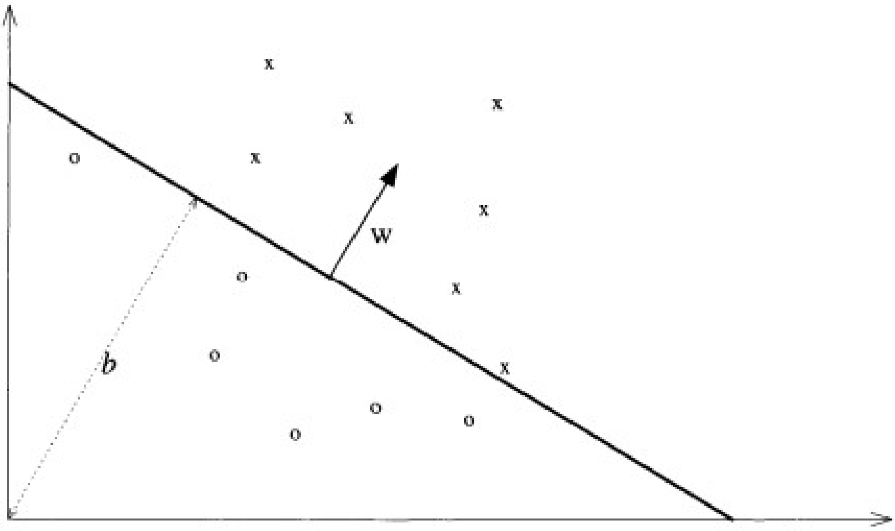
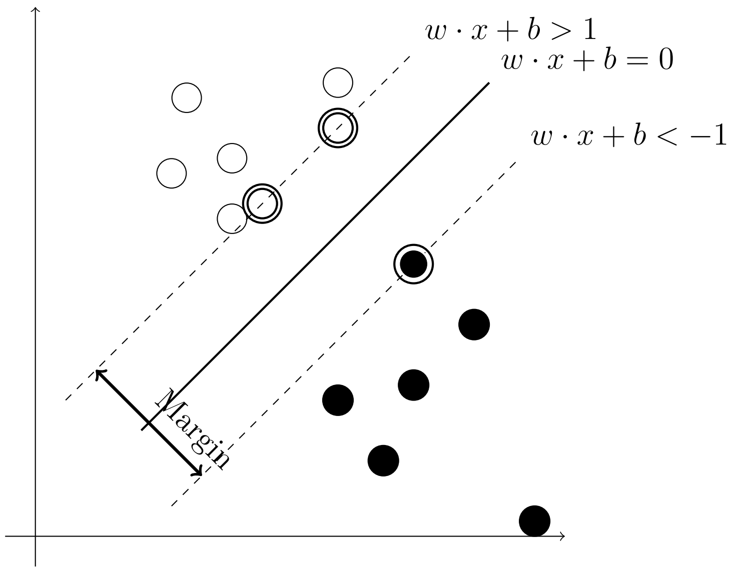


Figure 1: A hyperplane (w, b) showed in Cristianini and Shawe-Taylor [5] with a two-dimensional training dataset

do a multi-class classification. The idea of multi-class classification is separating the classes to a linear classification [18].

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 dimnensionality of the feature space is not relevant. This separation can be done in every kernel-induced feature space. [5]



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.

2.4 The threat model for attacker characteristics

In their paper, Doynikova et al. [6] show a formal attacker model with input data for experiments, the data handling process and describe the experiment that was executed. Doynikova et al. explain that the attacker models 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 three 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 last 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.

2.5 Relevant standards for risk measurement

This present thesis based the requirements of risk measurement of ISO/IEC 27004:2009. ISO/IEC 27004:2009 - Risk Measurement is a international security standard from the ISO 27000 [9] family which guides on continious basis evaluation methods. 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 [1]. The ISO/IEC 27004:2009 - Risk Measurement standard specifies what to be measured, when the measurement is needed and types of measurement [8]. Barabanov et al. [1] and Tarnes [16] 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.

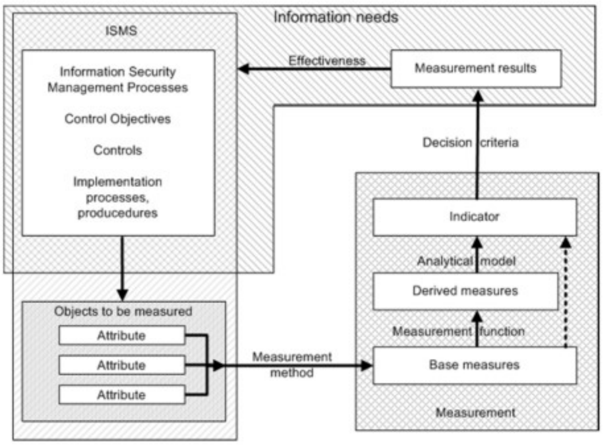


Figure 2: The information security measurement model [16]

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 3.4. The attributes in Figure 2 are the properties in the SMF.

2.6 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. [4] 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. [13] 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.

2.7 Security risks in context of Machine Learning

Security risks in context of ML must be derived from classic IT security risks in context to classic applications. Xiao et al. [19] evaluate the security risks in deep learning for common frameworks, for example TensorFlow. Xiao et al. uses 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. Xiao et. al work is only in context of deep learning e.g. only for neural networks.

Poisoning Attacks

Backdoor Attacks

Due to the rising amount of training data, human supervision to check trustworthiness is less possible. That exposes vulnerabilites in training datasets like backdoors. Backdoor attacks can cause farreaching consequences. Backdoored models are able to classify on most inference inputs. But it can cause targeted missclassifications or can decrease the accurarcy for inputs that the attackers choose as secret properties referring as backdoor trigger [7]. For example if the ML model classifies diseases with clinical pictures such as cancer most of the classifications have a good accuracy but then classifing a clinical picture with a certain conspicuity, that could potentially misclassify the right disease.

2.8 Risk assessment in context of Machine Learning

Risk assessment in context of ML is derived from classic IT security risk assessment. This subsection discusses paper from classic IT security risk assessment and abstract them to ML. Sendi et al. [14] 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 appraisementn perspective, 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 an 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 because for example in context of this thesis a manipulated trainings and testing dataset could lead a more extent missclassification while training and testing.

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 attacker effort. The attacker effort is measured by objective properties. These objective properties are the base of the risk indicators for the attacker effort explained in the following subsection. Objective properties

3.1 UML diagrams

3.2 Finding the attacker's effort

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

3.3 Characteristics of backdoor attacks

3.4 Risk indicators

The RMF measure risks by so called risk indicators. Properties, attributes and proposals are the basis for the risk indicators, among other things. Breier et al. in subsection 2.6 present proposals that are the approach for the proposals for the risk indicators.

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. [17] explain *PoisoningAttack-CleanLabelBackdoor* attacks. Gu et al. [7] explain *PoisoningAttackBackdoor* attacks. Gu et al. [7] 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 originial 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 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:

```
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 the different Python libraries for ML risk measurement.

4.4 Implementation of the visualization

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

4.5 Build in the risk indicatiors

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

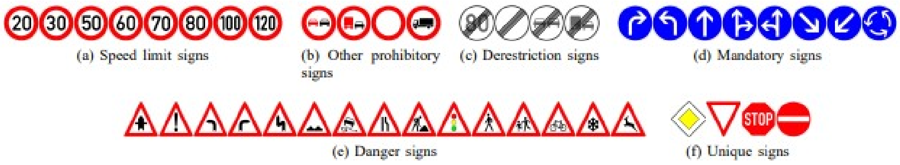
5 Evaluation

5.1 Case Study: Developing a SVM for traffic sign detection

For the case study scikit-learn [12] and for preparation of the dataset in Python OpenCV2 have different function to load and resize images [3]. In their work, Stallkamp et al. [15] 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.

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 3 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.



(a) Speed limit signs (b) Other prohibitory signs (c) Derestriction signs (d) Mandatory signs

(e) Danger signs (f) Unique signs

Figure 3: Labeled traffic signs [15]

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] Rostyslav Barabanov, Stewart Kowalski, and Louise Yngström. Information security metrics: State of the art: State of the art. 2011.

[2] Ekaba Ononse Bisong. *More Supervised Machine Learning Techniques with Scikit-learn*, page 287–308. Apress, 2019.

[3] G. Bradski. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*, 2000.

[4] Jakub Breier, Adrian Baldwin, Helen Balinsky, and Yang Liu. Risk management framework for machine learning security. *CoRR*, abs/2012.04884, 2020.

[5] Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, 2000.

[6] 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, CRiSIS 2020, Paris, France, November 4-6, 2020, Revised Selected Papers*, volume 12528 of *Lecture Notes in Computer Science*, pages 347–353. Springer, 2020.

[7] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *CoRR*, abs/1708.06733, 2017.

[8] 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.

[9] Ines Meriah and Latifa Ben Arfa Rabai. Comparative study of ontologies based ISO 27000 series security standards. In Elhadi M. Shakshuki, Ansar-Ul-Haque Yasar, and Haroon Malik, editors, *The 10th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (EUSPN 2019) / The 9th International Conference on Current and Future Trends of Information and Communication Technologies in Healthcare (ICTH-2019) / Affiliated Workshops, Coimbra, Portugal, November 4-7, 2019*, volume 160 of *Procedia Computer Science*, pages 85–92. Elsevier, 2019.

[10] 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.

[11] 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.

[12] 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.

[13] Paul Schwerdtner, Florens Greßner, Nikhil Kapoor, Felix Assion, René Sass, Wiebke Günther, Fabian Hüger, and Peter Schlicht. Risk assessment for machine learning models. *CoRR*, abs/2011.04328, 2020.

[14] Alireza Shameli Sendi, Rouzbeh Aghababaei-Barzegar, and Mohamed Cheriet. Taxonomy of information security risk assessment (ISRA). *Comput. Secur.*, 57:14–30, 2016.

[15] 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.

[16] Marte Tarnes. Information security metrics: An empirical study of current practice. *Specialization Project, Trondheim, 17th December*, 2012.

[17] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks. 2018.

[18] Angelos Tzotsos and Demetre Argialas. Support vector machine classification for object-based image analysis. In *Object-Based Image Analysis*, pages 663–677. Springer, 2008.

[19] 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 17, 2022 ....