22. Nachwuchswissenschaftler*innenkonferenz (NWK)

https://doi.org/10.52825/ocp.v2i.161

Publiziert: 15. Dez. 2022

Bedrohungsmodellierung im Machine Learning

Manuel Raddatz[1]

[1] Technische Hochschule Brandenburg, Fachbereich Wirtschaft

* Korrespondenz: manuel.raddatz@th-brandenburg.de

Kurzfassung. Aufgrund zunehmender Globalisierung, der technologischen Weiterentwicklung und dem Grad der Vernetzung steigt die Anzahl der Bedrohungen kontinuierlich an. Sicherheitsanforderungen spielen oft nur eine untergeordnete Rolle. Mit der neuen Fassung zum IT-Sicherheitsgesetz (Kurzform für "Gesetz zur Erhöhung der Sicherheit informationstechnischer Systeme") wurde dessen Geltungsbereich ausgedehnt und betroffene Unternehmen müssen Maßnahmen für die Steigerung der IT-Sicherheit ergreifen. Die Bedrohungsmodellierung ist ein strukturierter Prozess, der bereits in der sicheren Hard- und Software Anwendung findet. Sowohl Charakter der Angriffe als auch Zeitraum im Lebenszyklus unterscheiden sich zur traditionellen SW-Entwicklung. Ausgehend von der Machine Learning Struktur bietet dieser Beitrag einen Top-Down-Ansatz für die systemorientierte Perspektive der Bedrohungsmodellierung.

1 Einleitung

Der notwendige Aufwand für Angriffe auf IT-Systeme hat in den letzten Jahren abgenommen. Aus diesem Grund definieren verschiedene Normen und Standards Anforderungen an die verschiedenen Phasen des Lebenszyklus für eine sichere Produktentwicklung [1] und der Durchführung einer Risikobewertung [2]. Die Nutzung des maschinellen Lernens ist in den letzten Jahren sehr stark gestiegen. Die Fähigkeiten zu lernen, ist eine der wichtigsten Eigenschaften von intelligentem Verhalten. Lernprozesse umfassen den Erwerb von neuem Wissen und der Entwicklung von neuen kognitiven und motorischen Fähigkeiten. Im Allgemeinen definiert das Lernen die Organisation von neuem Wissen mit einer effektiven Darstellung und der Entdeckung neuer Zusammenhänge durch das Beobachten und Experimentieren [3]. Das maschinelle Lernen wird bereits erfolgreich in den Bereichen für das autonome Fahren, der Bilderkennung, der intelligenten Gesundheitsversorgung, der Sicherheit und Sprachverarbeitung eingesetzt. Parallel zur steigenden Verbreitung ist die Nutzung von Machine Learning einer zunehmenden Bedrohung ausgesetzt [4].

2 Bedrohungen im Machine Learning

Aufgrund des breiten Anwendungsspektrums von Machine Learning ist die Technologie sehr vielen unterschiedlichen Bedrohungen ausgesetzt, die sowohl beabsichtigt als auch unbeabsichtigt zu Sicherheitsproblemen führen können.

In [5] und [6] wurden eine Taxonomie für die Bestimmung von Angriffen auf Machine Learning Modelle definiert. Dementsprechend basiert die Klassifizierung eines Angriffs auf die Eigenschaften Influence (Einfluss), Security Violation (Verletzung der Sicherheitsziele) und Specificy (Spezifität). Die nachfolgende Tabelle zeigt einen Überblick der Bedrohungskategorien mit Einordnung in die Dimensionen Taxonomie und den Phasen des Machine Learning Lebenszyklus.

Tabelle 1: Angriffsformen im Machine Learning

Angriffsform	Kurzbeschreibung	Phase im Lebenszyklus	Taxonomie
Poisoning	Böswillige Veränderungen der Trainingsdaten [7].	Training und Test	Influence
Evasion	Täuschung des Machine Learning Modells durch minimale Veränderungen der Eingabedaten [8].	Test und Deployment	Influence, Security Violation und Specifity
Inversion	Rekonstruktion der nichtsichtbaren Eingabedaten [9] sowie die Ermittlung der genutzten Trainingsdaten [10].	Deployment	Security Violation
Extraction	Über Beobachten der Einschätzungen des Modells zu den Ein- und Ausgabedaten wird dieses rekonstruiert [11].	Test und Deploment	Security Violation

Analog zu artverwandten Disziplinen wie der Hard- und Softwareentwicklung unterliegen Machine Learning

Modelle einem Lebenszyklus. Aktuell existiert kein zentral definiertes Modell. Die spezifischen Angriffe der Bedrohungskategorien richten sich gegen unterschiedliche Phasen im Lebenszyklus eines Machine Learning Modells.

Abbildung 1 zeigt einen Entwurf für ein ML-Lebenszyklus-Modell, das auf den Vorgaben von [12], [13] und [14] basiert. Zusätzlich sind den Phasen die Bedrohungskategorien zugeordnet.

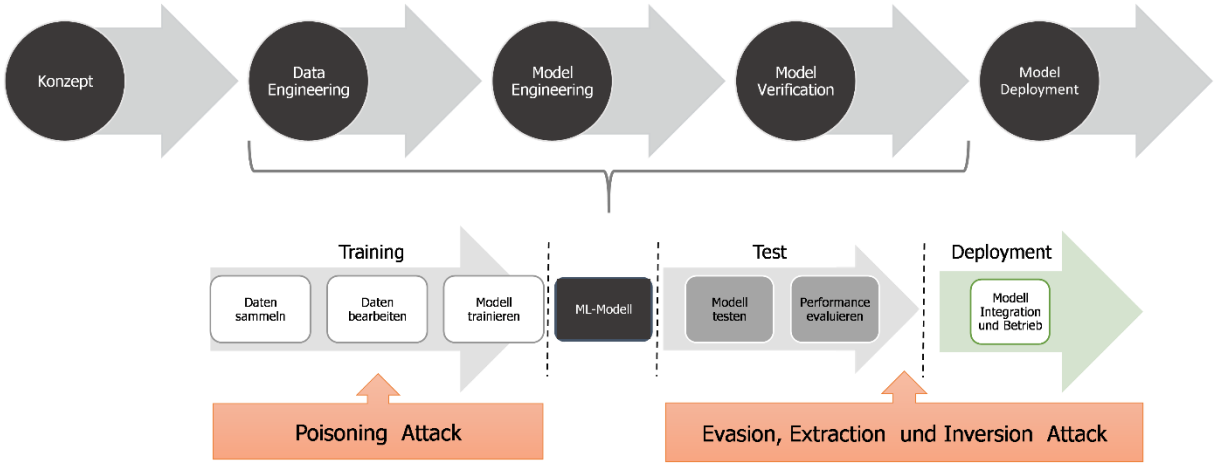

Abbildung 1: Bedrohungen für Machine Learning im Lebenszyklus vgl. [15]

3 Bedrohungsmodellierung

Sicherheitskonzepte und -modelle müssen im Rahmen eines Security by Design-Ansatz nachhaltig entwickelt und integriert werden. Im weitesten Sinn handelt es sich bei der Bedrohungsmodellierung um einen systematisch organisierten und methodisch strukturierten Prozess, der bereits im Rahmen von sicherer Software und Hardware Anwendung findet. [16]

Abbildung 2: Bedrohungen für Machine Learning im Lebenszyklus vgl. [17]

Unterschiedliche Perspektiven werden genutzt, um ein tiefes Verständnis für das Modell, den Betrieb aber auch den möglichen Angreifer zu erhalten [17]. In der *angreiferorientierten Perspektive* liegt der Fokus auf den/die möglichen Angreifer. Diese Betrachtungsweise entspricht dem natürlichsten Vorgang für die Bedrohungsmodellierung. Dabei stehen Fragen zur Motivation und zum Niveau im Mittelpunkt der Analyse. Bei dem Wissen zum Angreifer geht es vorrangig nicht um die Identifikation einer einzelnen Person, als vielmehr um die Erfassung der entsprechenden Angreifergruppe und dem daraus resultierenden Umfang, dem Können und der Motivation. Das Ergebnis der Analyse liefert weniger technischen Input für die Implementierung von Schutzmaßnahmen als vielmehr das Wissen zum weiteren Vorgehen sowie der Priorisierung von Gegenmaßnahmen. [18]

Die *wertorientierte Perspektive* bildet im Prozess die Phase zur Identifikation der zu schützenden Werte gemäß Abbildung 2 ab. Sie ist ein sehr wichtiger Einstieg in den Prozess der Bedrohungsmodellierung. Sie ist die Brücke zwischen dem Modell der Domäne und dem Erfassen der spezifischen Bedrohungen. Es werden sowohl die Anforderungen der Domäne als auch die der beteiligten Stakeholder berücksichtigt. Neben der Erfassung aller potentiellen Bedrohungen spielt die Effizienz eine wichtige Rolle. Ohne die richtigen und wichtigen Werte zu kennen, sind die nachgelagerten Phasen weniger effizient. [19]

Bei der *strukturorientierten Perspektive* wird das Machine Learning System in seine einzelnen Funktionsbereiche zerlegt und die Komponenten separat analysiert. Im Rahmen dieser Sichtweise wird das System auf eine Art und Weise modelliert, die es erlaubt, Bedrohungen bereits in der Entwurfsphase zu erfassen. In diesem Zusammenhang ist es ausreichend, dass nur die wesentlichen Komponenten modelliert werden, ohne ins Detail zu gehen. [18] Anhand der strukturorientierten Perspektive wird der iterative Charakter der Bedrohungsmodellierung deutlich. In kleinen, aufeinander abgestimmten Schritten werden die Sicherheitsvorgaben verfeinert. Die wichtigen Informationen im Rahmen der strukturorientierten Bedrohungsmodellierung sind Datenhaltung, Datenfluss und Datentransformation. Aus den genannten Bausteinen lassen sich die Vertrauensgrenzen für die Sicherheitsarchitektur ableiten. [17]

Für die strukturorientierte Bedrohungsmodellierung stehen unterschiedliche Notationen zur Auswahl. Eine Option bietet das Datenflussdiagramm (DFD). Es entstammt der Methode für die strukturierte Systemanalyse (SSA) und wurde von Gane und Sarson [20] in den 1970er Jahren entwickelt. Ein DFD ist ein Werkzeug für die Modellierung des Datenflusses. Sowohl in der klassischen Software als auch in Machine Learning Anwendungen bilden Daten den wichtigsten Baustein. Im Zusammenwirken mit den Methoden der Bedrohungsidentifikation bietet ein DFD eine sehr gute Möglichkeit Bedrohungen zu erfassen. [17]

Ein DFD zeigt den Datenfluss zwischen den Akteuren eines Systems. Dabei kann es sich um interne Komponenten oder externe Systeme handeln. Ihre Gemeinsamkeit ist die Datenverarbeitung. Die Modellierung bietet drei Hierarchiestufen für die Definition der Granularität der Modelle:

- DFD Level 0: In der ersten Hierarchiestufe werden die Datenflüsse über die Systemgrenzen hinweg beschrieben. Das Modell dieser Stufe wird als Kontextdiagramm bezeichnet [21].
- DFD Level 1: Die zweite Stufe dokumentiert die einzelnen Datenflüsse zwischen den internen und externen Systemkomponenten, die miteinander verbunden sind.
- DFD Level 2: In der dritten Stufe werden die Datenflüsse zwischen den internen und externen Systemkomponenten modelliert, die in den einzelnen Prozessschritten eines Anwendungsfalls miteinander verbunden sind.

Die modellierten Komponenten werden über die Hierarchiestufen hinweg zerlegt womit die einzelnen Datenflüsse detaillierter beschrieben werden.

3.1 Struktur des Machine Learning

Die Struktur des Lernens zeigt sich auf unterschiedliche Art und Weise. Sowohl Zeit, Raum als auch Lernform bieten Möglichkeiten für die Unterscheidung.

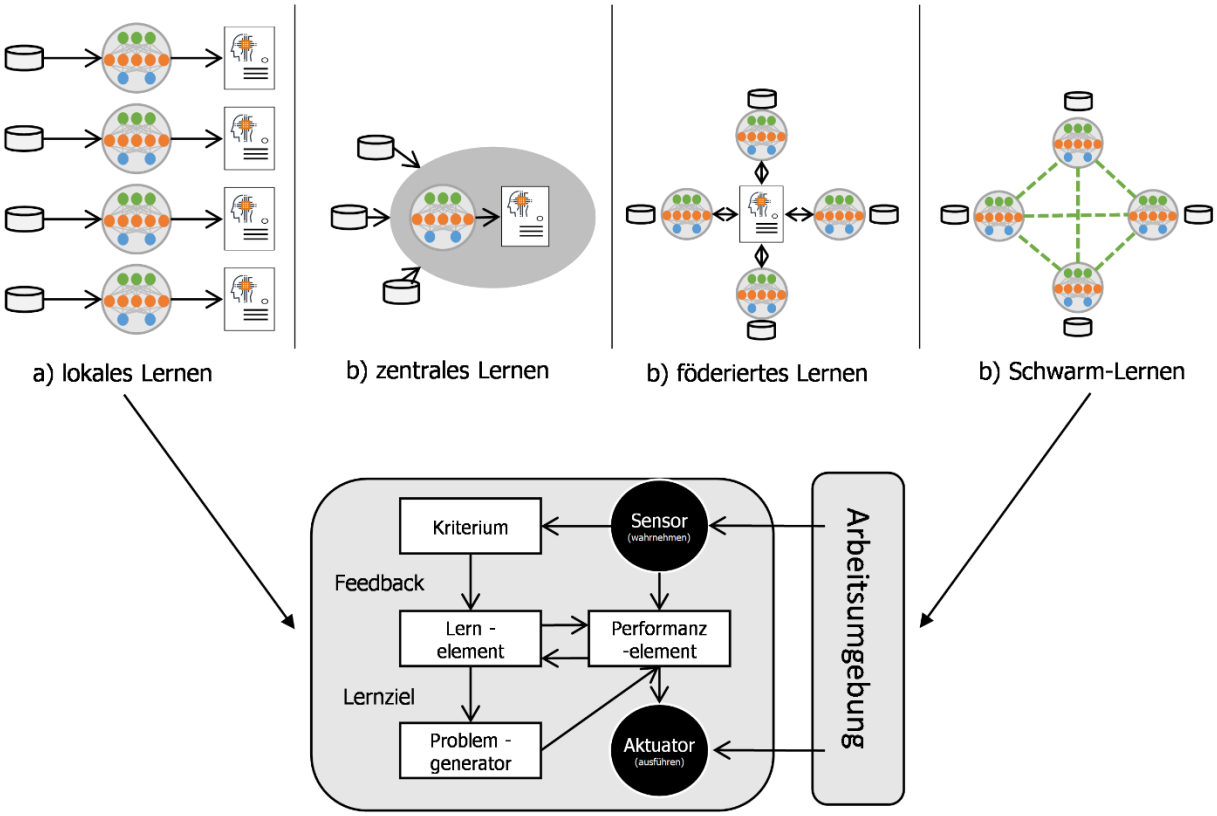

Abbildung 3: Machine Learning Struktur vgl. [22] und [23]

Abbildung 3 zeigt die hierarchische Sichtweise der Machine Learning Struktur. Die oberste Schicht bildet die Machine Learning Architektur. Die genutzte Architektur bestimmt den grundlegenden Aufbau der beteiligten Systeme und Agenten. Die zweite Schicht ist über den Aufbau des Machine Learning Agenten und seiner Arbeitsumgebung definiert. Sowohl Auswahl als auch Implementierung der Elemente des Agenten und der Komponenten der Arbeitsumgebung haben einen wesentlichen Einfluss auf das Gesamtsystem. Die Lernform bildet ein weiteres Strukturelement im Machine Learning. In dieser Stufe entscheiden Leistungs- und Lernelement des Agenten über das Lernproblem. Der vorhandene Feedbacktyp ist verantwortlich für das gegenwärtige Lernproblem des Agenten. Dabei spiegeln die unterschiedlichen Feedbacktypen die Lernformen wider. [22]

3.2 Strukturorientierte Bedrohungsmodellierung im Machine Learning

Die Abbildung 4 zeigt exemplarisch ein DFD. Die Architektur folgt dem zentralen Lernen. Ziel ist das Erlernen einer Regressionsfunktion.

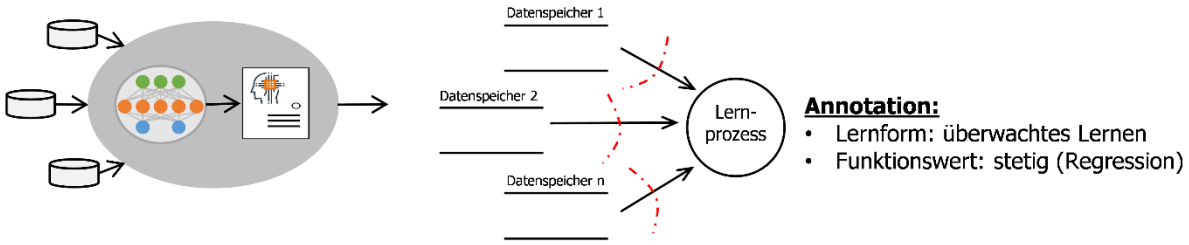

Abbildung 4: Machine Learning DFD

Über die Architektur mit den dislozierten Datenspeichern und dem zentralen Lernsystem, welches als Prozess modelliert ist, besteht in der Poisoning-Attacke ein mögliches Bedrohungsszenario. Über entsprechende Annotationen zur geplanten Umsetzung der Agentenfunktion können den DFD-Symbolen zusätzliche Informationen beigefügt werden, um über ein DFD-Modell die Struktur des maschinellen Lernens abzudecken. Im weiteren Prozessverlauf können die konkreten Gegenmaßnahmen automatisch bestimmt werden. In Anbetracht der zeitlichen Differenzierung von Bedrohungen im Lebenszyklus eines Machine Learning Modells ist es notwendig, dass die Modellierung für die unterschiedlichen Phasen durchgeführt wird, um das gesamte Bedrohungsspektrum abzudecken.

4 Zusammenfassung und Ausblick

Die Gründe für das Durchführen einer Bedrohungsmodellierung können sehr vielfältig sein. Hauptsächliches Ziel ist die Identifizierung von möglichst allen Bedrohungen für ein System, sowie die Erfassung und Planung von entsprechenden Gegenmaßnahmen, die die Auswirkungen von Angriffen verhindern bzw. reduzieren sollen. Die Bedrohungsmodellierung ist ein Teil des Security Engineering Prozess [24]. Hier nimmt sie eine gesonderte Stellung ein. Für die Umsetzung der Bedrohungsmodellierung existieren unterschiedliche Methoden, die in [25] und [26] zusammengefasst werden.

Aufgrund der vielen Ähnlichkeiten zwischen dem Machine Learning und der Softwareentwicklung liegt es nahe, dass das vorhandene Wissen und die Methoden zur Bedrohungsmodellierung für das Machine Learning angepasst werden. Methoden zur Bedrohungsidentifikation wie STRIDE [18] oder LINDDUN [27] nutzen die strukturorientierte Perspektive. Die Methoden enthalten eine feste Zuordnung der Bedrohungsformen für die einzelnen Elemente eines DFD, um den weiteren Prozess zu automatisieren. Ein ähnliches Vorgehen kann für die Bedrohungsmodellierung im Machine Learning eingeführt werden. Dafür ist es notwendig, den DFD-Symbolen den entsprechenden Bedrohungskategorien zuzuordnen. Für eine weitere Spezifizierung der Angriffe und Gegenmaßnahmen sind zusätzliche Annotationen für die einzelnen Symbole notwendig. Diese Symbole beschreiben die weiteren Hierarchiestufen der Machine Learning Struktur. Anhand der beigefügten Informationen und einer internen Threat Intelligence ist es möglich, dass die notwendigen Gegenmaßnahmen als zusätzliche Schritte im ML-Lebenszyklus automatisiert geplant werden.

Datenverfügbarkeit

Der Beitrag basiert nicht auf Daten, bis auf die genannten Quellen im Literaturverzeichnis.

Interessenskonflikte

Der Autor bestätigt, dass zum Beitrag keine Interessenkonflikte bestehen.

Literaturverzeichnis

1. „Sicherheit für industrielle Automatisierungssysteme," Geneva, 2019.
2. J. T. F. T. Initiative, „Guide for Conducting Risk Assessments," Washington, D.C., 2012.
3. R. S. Michalski, J. G. Carbonell und T. M. Mitchell, Machine Learning: An Artificial Intelligence Approach, 1 Hrsg., Los Altos: Morgan Kaufmann, 1983.
4. M. Xue, C. Yuan, H. Wu, Y. Zhang und W. Liu, „Machine learning security: Threats, countermeasures, and evaluations," IEEE Access, Bd. 8, p. 74720–74742, 2020.
5. L. Huang, A. D. Joseph, B. Nelson, B. I. P. Rubinstein und J. D. Tygar, „Adversarial machine learning," in Proceedings of the 4th ACM workshop on Security and artificial intelligence, 2011.
6. M. Barreno, B. Nelson, R. Sears, A. D. Joseph und J. D. Tygar, „Can Machine Learning Be Secure?," in Proceedings of the 2006 ACM Symposium on Information, computer and communications security, Taiwan, 2006.
7. B. Biggio, B. Nelson und P. Laskov, „Poisoning Attacks against Support Vector Machines," in 29th International Conference on Machine Learning (ICML 2012), Edinburgh, 2012.
8. B. Nelson, M. Barreno, F. J. Chi, A. D. Joseph, B. I. P. Rubinstein, U. Saini, C. Sutton, J. D. Tygar und K. Xia, „Exploiting machine learning to subvert your spam filter.," LEET, Bd. 8, p. 16–17, 2008.
9. M. Fredrikson, S. Jha und T. Ristenpart, „Model inversion attacks that exploit confidence information and basic countermeasures," in Proceedings of the 22nd ACM SIGSAC conference on computer and communications security, 2015.
10. R. Shokri, M. Stronati, C. Song und V. Shmatikov, „Membership inference attacks against machine learning models," in 2017 IEEE symposium on security and privacy (SP), 2017.
11. F. Tramèr, F. Zhang, A. Juels, M. K. Reiter und T. Ristenpart, „Stealing Machine Learning Models via Prediction {APIs}," in 25th USENIX security symposium (USENIX Security 16), 2016.
12. R. Ashmore, R. Calinescu und C. Paterson, „Assuring the machine learning ifecycle: Desiderata, methods, and challenges," ACM Computing Surveys (CSUR), Bd. 54, p. 1–39, 2021.
13. R. Garcia, V. Sreekanti, N. Yadwadkar, D. Crankshaw, J. E. Gonzalez und J. M. Hellerstein, „Context: The missing piece in the machine learning lifecycle," in KDD CMI Workshop, 2018.
14. R. Souza, L. Azevedo, V. Lourenço, E. Soares, R. Thiago, R. Brandão, D. Civitarese, E. Brazil, M. Moreno, P. Valduriez, M. Mattoso, R. Cerqueira und M. A. S. Netto, „Provenance Data in the Machine Learning Lifecycle in Computational Science and Engineering," in 2019 IEEE/ACM Workflows in Support of Large-Scale Science (WORKS), 2019.
15. Q. Liu, P. Li, W. Zhao, W. Cai, S. Yu und V. C. M. Leung, „A survey on security threats and defensive techniques of machine learning: A data driven view," IEEE access, Bd. 6, p. 12103–12117, 2018.
16. A. Selzer, H. Schöning, M. Laabs, S. Dukanovic und T. Henkel, IT-Sicherheit in Industrie 4.0 - Mit Bedrohungen und Risiken umgehen, 1 Hrsg., Stuttgart: Kohlhammer Verlag, 2020.
17. S. Paulus, Basiswissen Sichere Software - Aus- und Weiterbildung zum ISSECO Certified Professionell for Secure Software Engineering, 1 Hrsg., Heidelberg: dpunkt.verlag, 2012.
18. A. Shostack, Threat Modeling - Designing for Security, New York: John Wiley & Sons, 2014.
19. N. Messe, V. Chiprianov, N. Belloir, J. El-Hachem, R. Fleurquin und S. Sadou, „Asset-Oriented Threat Modeling," in 2020 IEEE 19th International Conference on Trust, Security and Privacy in Computing and Communications (TrustCom), 2020.
20. C. Gane und T. Sarson, Structured Systems Analysis - Tools and Techniques, 1 Hrsg., New York: Prentice-Hall, 1979.

21. R. Ibrahim und S. Y. Yen, "An automatic tool for checking consistency between Data Flow Diagrams (DFDs)," World Academy of Science, Engineering and Technology, Bd. 69, p. 2010, 2010.
22. S. J. Russell und P. Norvig, Artificial Intelligence - A Modern Approach, 3 Hrsg., London: Prentice Hall, 2010.
23. S. Warnat-Herresthal, H. Schultze, K. L. Shastry, S. Manamohan, S. Mukherjee, V. Garg, R. Sarveswara, K. Händler, P. Pickkers, N. A. Aziz und others, "Swarm learning for decentralized and confidential clinical machine learning," Nature, Bd. 594, p. 265-270, 2021.
24. "Information technology - Security techniques - Systems Security Engineering - Capability Maturity Model," Geneva, 2008.
25. K. Tuma, G. Calikli und R. Scandariato, "Threat analysis of software systems: A systematic literature review," Journal of Systems and Software, Bd. 144, pp. 275-294, 2018.
26. W. Xiong und R. Lagerström, „Threat modeling – A systematic literature review," Computers & Security, Bd. 84, pp. 53-69, 2019.
27. M. Deng, K. Wuyts, R. Scandariato, B. Preneel und W. Joosen, „A privacy threat analysis framework: supporting the elicitation and fulfillment of privacy requirements," Requirements Engineering, Bd. 16, p. 3–32, 2011.