Hochschule der Medien Stuttgart
Sommersemester 2022
Aktuelle Themen (AM3)
Verena Eichinger (ve018)
Matrikel-Nr.: 40833

Sicherheit und Privatsphäre von KI

EINLEITUNG

KI ist ein Thema, das heutzutage den Weg aus wissenschaftlichen Fachkreisen in den Alltag und die Medien gefunden hat. Beinahe täglich berichten die Nachrichten über neueste Anwendungen und Entwicklungen im Bereich der KI. Hierbei werden auch die Sicherheit und Privatsphäre von KI-Anwendungen regelmäßig thematisiert. Im Rahmen dieser Arbeit wird genau dieser Bereich näher betrachtet, jedoch mit der Eingrenzung auf Machine Learning (ML) Anwendungen. Machine Learning ist ein Teilbereich der KI, in dem Algorithmen aus vorgegebenen Daten lernen, und darauf basierend Vorhersagen oder Entscheidungen treffen. Die genutzten Daten werden vom Menschen aufbereitet und die Algorithmen erlernen daraus im Training Muster, die sie zur Vorhersage nutzen. [1] Die Ausgabe, die ein solcher Algorithmus nach dem Training liefert, wird als Machine Learning Modell bezeichnet. Für den Betrieb der Anwendung werden dem Modell neue Daten als Input geliefert, um darauf basierend Vorhersagen zu erhalten. [2]

Mit der zunehmenden Verbreitung derartiger Anwendung hat sich „Machine Learning as a Service“ (MLaaS) entwickelt. Unter diese Bezeichnung fallen Cloud-basierte Plattformen, die Tools für ML-Anwendungen anbieten. Dabei werden Services wie die Vorverarbeitung der Daten, das Training eines Machine Learning Modells oder das Deployment des Modells angeboten. Anbieter für MLaaS sind beispielsweise Amazon, Google oder Microsoft. [3] Auch hier stellt sich die Frage zur Sicherheit und der Privatsphäre. Machine Learning Anwendungen benötigen meist große Datenmengen zum Training, wobei diese in vielen Fällen privat bleiben sollen. Einige Firmen bieten die Nutzung ihrer ML-Anwendungen gegen eine Gebühr an, hier spielt zusätzlich die Geheimhaltung des Modells eine große Rolle, da dieses den wirtschaftlichen Wert eines solchen Dienstes darstellt. Besonders da Machine Learning mittlerweile in allen Lebensbereichen, von der Finanzwelt bis hin zum Gesundheitswesen, Einzug gehalten hat, und dadurch sensible Daten in die Modelle einfließen, gilt es die Sicherheit und Privatsphäre derartiger Anwendungen ernst zu nehmen. [4] Gelingt es einem Angreifer aus der Nutzung einer Anwendung Informationen über das Modell oder die genutzten Daten zu gewinnen, so stellt das ein großes Risiko dar. Der Schutz der gespeicherten Daten an sich kann in die klassische IT-Sicherheit eingeordnet werden. Das Machine Learning bietet allerdings auch während dem Training und dem Betrieb der Modelle Angriffspunkte. Die nachfolgend erläuterten Attacken setzen an diesen Bereichen an und sollen aufzeigen, dass das Risiko von ML-Anwendungen über die Speicherung der Daten hinaus geht.

DATEN IM MACHINE LEARNING

Zunächst wird ein Überblick über die vorliegenden Daten im Machine Learning gegeben, sodass die Ansätze der unterschiedlichen Attacken leichter nachvollzogen werden können.

Da Machine Learning Modelle basierend auf den vorliegenden Daten (Trainingsdaten) erstellt werden, liegen folglich zwei wesentliche Bestandteile vor, das Modell selbst und die genutzten Trainingsdaten. [5]

Die Trainingsdaten werden als Input für das Training des Modells genutzt. Es handelt sich meist um große Datensätze, die als

Beispieldaten für den Anwendungsfall dienen. [5] Beispielsweise werden bisherige Einkäufe auf einer Online-Plattform dokumentiert, sodass für jeden Nutzer die gekauften Artikel mit Preis und Kaufdatum bekannt sind. Dieser Datensatz kann zum Training eines Modells genutzt werden, das den Kunden für sie interessante Artikel vorschlägt. Die Trainingsdaten liegen in der Regel den Firmen vor, die diese für ihre Anwendung sammeln. Dort werden sie zum Training genutzt und sind von entsprechendem Wert für die Eigentümer. Zu den Trainingsdaten zählen neben den Daten an sich auch statistische Eigenschaften der Daten oder die Zugehörigkeit zu den Daten. Letzteres ist die Information darüber, ob ein bestimmter Datensatz zum Training verwendet wird oder nicht. [4]

Das fertige Modell, das durch das Training mit den Trainingsdaten erzeugt wird, wird durch den Modell-Typ und seine Parameter definiert. Dabei werden anhand der Trainingsdaten Muster gelernt und nicht die konkreten Trainingsdaten. Diese sind im Modell nicht mehr als solche enthalten. Der Modell-Typ sowie ein Teil der Parameter wird in der Regel vom Entwickler vorgegeben. Die restlichen Parameter werden während des Trainings anhand der Trainingsdaten so optimiert, dass das resultierende Modell seine Aufgabe möglichst gut erfüllt. [5, 6] Wird eine Anwendung als MLaaS angeboten, so erfolgt der Zugriff meist über eine API, über die Anfragen gestellt werden können und anschließend ein Ergebnis eingesehen werden kann. Entsprechend sind die Informationen über das Modell und seine Parameter für den Nutzer nicht ersichtlich. [7, 4]

Ist das Training mit den Trainingsdaten abgeschlossen und ein fertiges Modell erstellt, so kann dieses in Betrieb genommen werden. Dabei werden neue Input-Daten als Eingabe geliefert, um durch das Modell Vorhersagen oder Entscheidungen zu treffen. Zusätzlich zum Modell und den Trainingsdaten liegen also während des Betriebs neue Input-Daten vor. [4]

Bei Angriffen auf ML-Anwendungen ist es von Bedeutung zu definieren, auf welche Informationen ein Angreifer Zugriff hat. Man unterscheidet dabei zwischen zwei Varianten. Hat ein Angreifer Zugang zum Modell und dessen Parametern handelt es sich um „white-box access“. Die zweite Möglichkeit ist der „black-box access“, hier hat ein Angreifer ausschließlich die Möglichkeiten eines Endnutzers und kann über ein Interface Anfragen an das Modell stellen, sowie die Ergebnisse abgreifen. [4]

PROBLEME FÜR PRIVATSPHÄRE UND SICHERHEIT

Die hier betrachteten Angriffe auf ML-Anwendungen zielen auf die Vertraulichkeit und Privatsphäre des Modells und der Trainingsdaten ab. Es werden also Informationen über die zuvor erläuterten Daten gewonnen. [8]

Für das Modell bedeutet dies konkret, dass mit den Informationen eine Kopie erstellt werden kann. Das birgt für Firmen, die Modelle als Service anbieten, das Risiko, dass die Gebühren für die Nutzung durch das Generieren einer solchen Kopie umgangen werden können. Das Resultat sind entsprechend der Verlust der Daten sowie wirtschaftliche Einbußen. [8]

Gelingt es einem Angreifer aus der Anwendung Informationen über die Trainingsdaten zu gewinnen, so besteht die Gefahr eines Daten Leaks. Da es sich bei den Trainingsdaten um sensible persönliche Daten handeln kann, wie im Falle einer Anwendung aus dem medizinischen Bereich, ist der Diebstahl oder Leak der Trainingsdaten ein großes Risiko für die Privatsphäre der Betroffenen. [8]

AKTUELLE ANGRIFFSMÖGLICHKEITEN AUF KI

Sowohl während des Trainings-Prozesses, als auch während dem Betrieb der Anwendung können Angriffe erfolgen, die auf die Informationen über das Modell selbst und seine Trainingsdaten abzielen. Neben dem Informationsgewinn aus der Anwendung gibt es auch Attacken, die das Ziel verfolgen die Ausgabe des Modells zu manipulieren oder den Betrieb des Modells durch Unzuverlässigkeit und Inkonsistenz negativ zu beeinflussen [8]. Derartige Attacken werden an dieser Stelle außen vor gelassen, sodass der Informationsgewinn aus den Modellen im Vordergrund steht. Dazu werden zunächst Angriffe erläutert, die auf das Modell und seine Parameter abzielen, anschließend werden Attacken betrachtet, die die Trainingsdaten in Erfahrung bringen möchten.

Angriffe mit dem Modell als Ziel

Kann ein Angreifer die Daten über ein Modell stehlen, so kann er durch die Erzeugung einer Kopie Kosten umgehen oder mit Hilfe dieser Informationen eigenen Produkten einen Vorteil gegenüber der Konkurrenz verschaffen. Zwei Varianten von Attacken, die dies ermöglichen werden nachfolgend erläutert.

Berechnung der Modell-Parameter

Der erste Ansatz verfolgt das Ziel die Modell-Parameter durch Berechnung zu bestimmen. Der Angriff erfolgt lediglich über „black-box access“ auf das Modell. Tramer et al. zeigen, dass es für mehrere Modell-Typen möglich ist eine sehr gute Kopie zu erstellen, wenn eine Anwendung als MLaaS Anfragen entgegen nimmt und das Ergebnis zurückliefert. Dabei ist es wichtig, dass neben dem Ergebnis auch ein Wahrscheinlichkeitswert für dieses angegeben wird. [9] Einer solchen Anwendung werden Anfragen gestellt und diese zusammen mit den erhaltenen Ergebnissen dokumentiert. Diese Paare von Input und Output werden als Gleichungen betrachtet, mit denen man die unbekannten Modell-Parameter berechnen kann. [8] Beschränkt ist dieser Angriff durch die gelieferten Wahrscheinlichkeitswerte für die Ergebnisse. Fehlen diese, ist eine derartige Attacke zwar noch möglich, jedoch mit einem sehr hohen Aufwand verbunden. [9] Tramer et al. fokussieren sich dabei auf die Parameter des Modells, die im Training gelernt werden. Weiterführende Arbeiten zeigen, dass auch die Parameter, die von Entwicklern festgelegt werden, ermittelt werden können. [7]

Erzeugung eigener Trainingsdaten

Im zweiten Angriff wird nicht das Ziel verfolgt die Modell-Parameter herauszufinden. Stattdessen soll ein eigenes Modell erzeugt werden, dass der attackierten Anwendung möglichst nahe kommt. Entsprechend liegt kein direkter Informationsgewinn über ein KI-Modell vor, jedoch soll das Ergebnis eine Kopie der Anwendung darstellen, so als würde das Modell gestohlen. Auch in diesem Fall liegt ein „black-box access“ zum Modell vor.

Um eine Modell-Kopie zu erstellen reicht es aus, einer Anwendung, die als Service Anfragen entgegen nimmt, eine Reihe von Eingaben zu liefern und diese mit den zurückgeliefert Ausgaben zu speichern. Die so gesammelten Daten bilden einen Datensatz, der anschließend zum Training eines neuen Modells genutzt werden kann. Ein Angreifer generiert so einen eigenen Trainingsdatensatz, mit dem er eine eigene Version des Modells trainiert. [10]

Angriffe mit den Trainingsdaten oder Input-Daten als Ziel

Da Machine Learning mittlerweile in allen Bereichen eingesetzt wird, kommen auch personenbezogene sensible Daten als Trainingsdaten und Input-Daten zum Einsatz. Dies ist beispielsweise bei Anwendungen in der Medizin der Fall. Derartige Daten können von großem Interesse für Angreifer sein. Um

Informationen über die Trainingsdaten oder Input-Daten zu gewinnen gibt es verschiedene Ansätze. Mehrere Angriffe aus diesem Bereich werden nachfolgend betrachtet.

Model Inversion

Bei der „Model Inversion“ erfolgt eine Umkehrung des Modells, die es möglich macht resultierende Modell-Ausgaben zu beobachten und daraus Trainingsdaten zu extrahieren. [11]

Abbildung 1: Rekonstruktion eines Bildes aus einer Gesichtserkennungs-Anwendung (links) und das zugehörige Bild aus den Trainingsdaten (rechts) [11]

Dieser Angriff ist dadurch limitiert, dass nicht ein bestimmter Datensatz bekannt wird, sondern eine Repräsentation des Durchschnitts aller Datensätze einer Klasse. Liegt jedoch eine Anwendung vor, bei der jeder Datensatz zu einer eigenen Klasse gehört, können spezifische Informationen gewonnen werden. [8] Fredrikson et al. testen einen solchen Angriff auf eine Gesichtserkennungs-Anwendung. Diese liefert zusätzlich zur Klasse einen Wahrscheinlichkeitswert für diese als Ergebnis. Zusätzlich zum Zugang zum Modell als „black-box access“ wird der Name einer Person benötigt. Damit rekonstruieren sie das Gesicht der Person und damit den Trainingsdatensatz. Abbildung 1 zeigt eine solche Rekonstruktion und die deutliche Ähnlichkeit zum Original aus den Trainingsdaten. [11]

Nutzung einer KI

Diese Art des Angriffs wird auch „Shadow Model Attack“ genannt. Dabei trainiert der Angreifer ein eigenes Machine Learning Modell, um Informationen aus einem anderen Modell zu gewinnen. [4]

Ateniese et al. wenden diese Methode auf eine Anwendung zur Spracherkennung an. In diesem Fall ist es das Ziel, statistische Information über die Trainingsdaten zu gewinnen statt konkrete Datensätze zu rekonstruieren. Mit ihrer Methode erhalten sie Informationen über die zum Training der Spracherkennung genutzte Akzente. Derartige Informationen können beispielsweise aufdecken, warum bestimmte Modelle bessere Ergebnisse liefern als andere. [12]

Manipulation des Trainings

Bei dieser Methode beeinflusst ein Angreifer das Training des Modells, sodass dieses später Informationen über die Trainingsdaten preisgibt. Dabei sollen einzelne Datensätze rekonstruiert werden, indem sie beim Training im Modell codiert werden. [4]

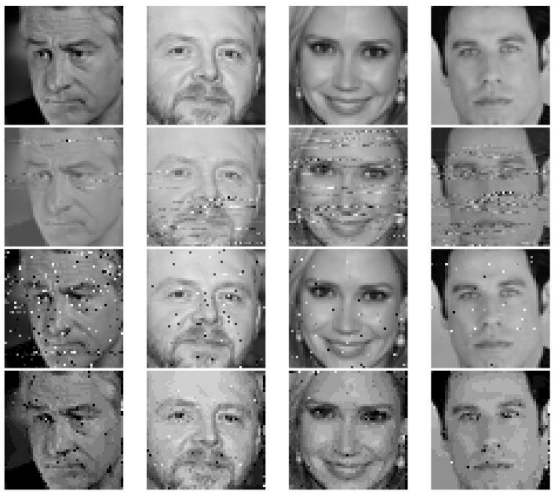

Abbildung 2: Bilder aus den Trainingsdaten (Zeile 1), Rekonstruktion mit „black-box access“ zum Modell (Zeile 2) Rekonstruktion mit „white-box access“ zum Modell (Zeile 3 und 4)

Song et al. Zeigen diese Methode am Beispiel eines bösartigen Anbieters für Machine

Learning. Ein auf das Training spezialisierter Anbieter hat keinen direkten Zugang zu den Trainingsdaten oder dem Training, jedoch zum fertigen Modell. Durch eine Manipulation des angebotenen Trainingsprozesses werden Informationen im Modell codiert. So können diese nach dem Training wieder ausgelesen werden. Es wurden sowohl Methoden für den „white-box access“ als auch für den „black-box access“ angewandt. Für den Fall dass ein Zugriff auf die Modell-Parameter möglich ist, werden die Trainingsdaten während des Trainings in den Parametern codiert, um später daraus ausgelesen zu werden. Für den „black-box access“ wird der Input während des Trainings durch synthetische Inputs manipuliert, durch die die Informationen codiert werden. [4] Die Ergebnisse des manipulierten Modells unterscheiden sich während dem Betrieb nicht von Ergebnissen eines nicht-manipulierten Modells. Abbildung 2 zeigt das Ergebnis der Attacke auf ein Modell, das zur Klassifikation der Geschlechter auf dem „FaceScrub“ Datensatz trainiert wurde. [13]

Identifizieren anonymisierter Daten

Dieser Angriff unterscheidet sich von den vorherigen Methoden, da in diesem Fall die Trainingsdaten anonymisiert öffentlich zugänglich sind. Dies ist beispielsweise bei Wettbewerben für die Erstellung eines Modells der Fall. Ein solcher Zugang zu den Daten ist dann problematisch, wenn Angreifer die Anonymisierung aufheben können. [14]

Es besteht die Möglichkeit die Einträge derartiger Datensätze einzelnen Personen zuzuordnen. Mit den dadurch erhaltenen Daten über Einzelpersonen können weitere Informationen abgeleitet werden. Durch anonymisierte Trainingsdatensätze besteht folglich ein Risiko für die Privatsphäre der Betroffenen. [14]

Ein Beispiel für einen derartigen Angriff zeigen Narayanan und Shmatikov anhand des Netflix Datensatzes von 500.000 Nutzern der Plattform. Sie identifizieren mithilfe weniger Informationen über einen Nutzer den jeweiligen Eintrag des Datensatzes. Mit den erhaltenen Daten ermitteln Sie zudem die mutmaßliche politischen Einstellung des Betroffenen. Für die Identifizierung benötigen sie Informationen zu gesehenen Filmen mit einer groben Datumsangabe, die angibt wann ein Film gesehen wurde. Mit den Daten zu acht Filmen eines Nutzers beträgt die Wahrscheinlichkeit den Datensatz aus den anonymisierten Daten zuordnen zu können 99%. Dabei dürfen zwei der acht Filme falsch sein und Datumsangaben um bis zu 14 Tage abweichen. Mit weniger Informationen sinkt die Wahrscheinlichkeit der Zuordnung. [15]

Seitenkanal-Attacken

Seitenkanal-Attacken nutzen physische Besonderheiten aus, die durch die Implementierung des Modells auftreten. Diese Besonderheiten können im Zeitverbrauch , im Stromverbrauch, im elektromagnetischen Feld oder der Akustik liegen. [16] Die genannten Eigenschaften werden bei einem solchen Angriff beobachtet, um daraus Informationen abzuleiten.

Abbildung 3: Ausschnitt des MNIST Datensatzes (Zeile 1), Trennung von Hintergrund und Vordergrund durch die Seitenkanal-Attacke (Zeile 2), pixelweise Rekonstruktion durch die Seitenkanal-Attacke (Zeile 3) [17]

Wei et al. rekonstruieren mit einer solchen Attacke Input Bilder eines Convolutional Neural Networks (CNN) indem sie den Stromverbrauch eines CNN-Accelerators während des Betriebs messen. [4] Ein derartiger Accelerator implementiert die Berechnung innerhalb des CNNs. In diesem Fall werden nicht die Trainingsdaten, sondern die

Input-Daten beim Betrieb des Modells rekonstruiert. Dabei liegt „black-box access“ zum Modell vor, die Parameter des Modells sind also nicht bekannt. Aus dem Stromverbrauch können Pixel mit ähnlichen Werten identifiziert werden. Diese dienen zur Bestimmung von Hintergrund und Vordergrund und der Einteilung des Bildes in diese zwei Bestandteile. Optimale Bilder für dieses Verfahren zeigen eine Form auf schwarzem Hintergrund, daher wird als Beispiel der MNIST Datensatz verwendet. Im Falle echter Anwendungen trifft dies aber auch auf Ultraschall- oder Radiographie-Bilder zu.

Abbildung 4: Mammographie-Bild (links), Trennung von Hintergrund und Vordergrund durch die Seitenkanal-Attacke (mitte), pixelweise Rekonstruktion durch die Seitenkanal-Attacke (rechts) [17]

Als weiterer Schritt wird eine pixelweise Rekonstruktion der Bilder angestrebt. Dazu ist für den Angreifer jedoch vor der eigentlichen Attacke Zugang zur Hardware nötig, da hierfür mehr Informationen über das Verhalten des Stromverbrauchs nötig sind. Abbildung 3 zeigt die Resultate auf dem MNIST Datensatz. Die Trennung von Hintergrund und Vordergrund sowie die Rekonstruktion ist auf diesen Bildern bereits sehr erfolgreich. Abbildung 4 zeigt die Anwendung auf Mammographie Bilder. An diesen erkennt man, dass die pixelweise Rekonstruktion bei derartig komplexen Bildern noch nicht funktioniert. [17]

FAZIT

Im Zeitalter von Social Media geben die meisten Menschen online viel über sich Preis, ohne weiter darüber nachzudenken. Viele Lebensbereiche finden zumindest teilweise auch digital statt. Jeder online Einkauf, jeder Like, jede gebuchte Zugfahrt oder sogar Arztbesuche resultieren in digital gespeicherten Daten. Diese Daten werden von verschiedensten Firmen und Institutionen gesammelt um daraus Vorteile zu ziehen. Ein Risiko für die Sicherheit und Privatsphäre der Daten besteht allerdings nicht nur durch das Sammeln und Speichern. Diese Punkte abzusichern ist zwar auch ein wesentlicher Aspekt, jedoch endet die Gefahr an dieser Stelle nicht. KI und Machine Learning finden zunehmend Anwendung im täglichen Leben und allen dazugehörigen Bereichen. Das bedeutet auch, dass Daten aus allen Bereichen in diese Anwendungen einfließen. Oft geht man davon aus, dass die eigenen Daten in den Anwendungen sicher sind, oder ohnehin keine Aussagekraft besitzen, doch der heutige Stand der Technik zeigt uns klar, wie falsch diese Annahmen sind. Bereits aus wenigen Daten leiten Machine Learning Anwendungen sensible Informationen wie die politische Gesinnung oder den Gesundheitszustand von Einzelpersonen ab. Dies macht es umso kritischer, dass wie zuvor aufgezeigt Informationen über das Modell selbst, über die Trainingsdaten oder die Daten während des Betriebs aus ML-Anwendungen gewonnen werden können.

Die hier gezeigten Verfahren stammen aus den letzten zehn Jahren und stellen damit vermutlich nur den Anfang eines Problems für die Privatsphäre und Sicherheit dar. Diese zu gewährleisten wird in Zukunft eine große Herausforderung darstellen, besonders, da sich die Angreifer und Angriffsmöglichkeiten genau so schnell weiterentwickeln wie die Technik selbst.

LITERATURVERZEICHNIS

[1] R. Ceron. „AI, machine learning and deep learning: What's the difference? - Servers & Storage". Servers & Storage. https://www.ibm.com/blogs/systems/ai-machine-learning-and-deep-learning-whats-the-difference/ (Zugriff am 1. Oktober 2022).

[2] J. Brownlee. „Difference Between Algorithm and Model in Machine Learning". Machine Learning Mastery. https://machinelearningmastery.com/difference-between-algorithm-and-model-in-machine-learning/ (Zugriff am 1. Oktober 2022).

[3] E. Onose. „Machine Learning as a Service: What It Is, When to Use It and What Are the Best Tools Out There - neptune.ai". neptune.ai. https://neptune.ai/blog/machine-learning-as-a-service-what-it-is-when-to-use-it-and-what-are-the-best-tools-out-there (Zugriff am 1. Oktober 2022).

[4] B. Liu, M. Ding, S. Shaham, W. Rahayu, F. Farokhi und Z. Lin, „When Machine Learning Meets Privacy", *ACM Computing Surveys*, Bd. 54, Nr. 2, S. 1–36, April 2021. Zugriff am: 1. Oktober 2022. [Online]. Verfügbar unter: https://doi.org/10.1145/3436755

[5] E. Alpaydin, *Introduction to machine learning*, 2. Aufl. Cambridge, MA: MIT Press, 2010.

[6] K. Nyuytiymbiy. „Parameters and Hyperparameters in Machine Learning and Deep Learning". Medium. https://towardsdatascience.com/parameters-and-hyperparameters-aa609601a9ac (Zugriff am 1. Oktober 2022).

[7] E. De Cristofaro, "An Overview of Privacy in Machine Learning", *ArXiv, abs/2005.08679*, 2020. Zugriff am: 1. Oktober 2022. [Online]. Verfügbar unter: https://arxiv.org/abs/2005.08679

[8] N. Papernot, P. Mcdaniel, A. Sinha und M. P. Wellman "Towards the Science of Security and Privacy in Machine Learning", *ArXiv, abs/1611.03814*, 2016. Zugriff am: 1. Oktober 2022. [Online]. Verfügbar unter: https://arxiv.org/abs/1611.03814

[9] F. Tramèr, F.Zhang, A. Juels, M. K. Reiter und T. Ristenpart, "Stealing Machine Learning Models via Prediction APIs", *ArXiv, abs/1609.02943*, 2016. Zugriff am: 1. Oktober 2022. [Online]. Verfügbar unter: https://doi.org/10.48550/arXiv.1609.02943

[10] T. Orekondy, B. Schiele und M. Fritz, „Knockoff Nets: Stealing Functionality of Black-Box Models", in *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Long Beach, CA, USA, 15.–20. Juni 2019. IEEE, 2019. Zugriff am: 1. Oktober 2022. [Online]. Verfügbar unter: https://doi.org/10.1109/cvpr.2019.00509

[11] M. Fredrikson, S. Jha und T. Ristenpart, „Model Inversion Attacks that Exploit Confidence Information and Basic Countermeasures", in *CCS'15: The 22nd ACM Conference on Computer and Communications Security*, Denver Colorado USA. New York, NY, USA: ACM, 2015. Zugriff am: 4. Oktober 2022. [Online]. Verfügbar unter: https://doi.org/10.1145/2810103.2813677

[12] G. Ateniese, L. V. Mancini, A. Spognardi, A. Villani, D. Vitali und G. Felici, „Hacking smart machines with smarter ones: How to extract meaningful data from machine learning classifiers", *International Journal of Security and Networks*, Bd. 10, Nr. 3, S. 137, 2015. Zugriff am: 1. Oktober 2022. [Online]. Verfügbar unter: https://doi.org/10.1504/ijsn.2015.071829

[13] C. Song, T. Ristenpart und V. Shmatikov, „Machine Learning Models that Remember Too Much“, in *CCS '17: 2017 ACM SIGSAC Conference on Computer and Communications Security*, Dallas Texas USA. New York, NY, USA: ACM, 2017. Zugriff am: 4. Oktober 2022. [Online]. Verfügbar unter: https://doi.org/10.1145/3133956.3134077

[14] M. Al-Rubaie und J. M. Chang, „Privacy-Preserving Machine Learning: Threats and Solutions“, *IEEE Security & Privacy*, Bd. 17, Nr. 2, S. 49–58, März 2019. Zugriff am: 4. Oktober 2022. [Online]. Verfügbar unter: https://doi.org/10.1109/msec.2018.2888775

[15] A. Narayanan und V. Shmatikov, „Robust De-anonymization of Large Sparse Datasets“, in *2008 IEEE Symposium on Security and Privacy (sp 2008)*, Oakland, CA, USA, 18.–22. Mai 2008. IEEE, 2008. Zugriff am: 4. Oktober 2022. [Online]. Verfügbar unter: https://doi.org/10.1109/sp.2008.33

[16] „Side-Channel Attack - Glossary | CSRC“. NIST Computer Security Resource Center | CSRC. https://csrc.nist.gov/glossary/term/side_channel_attack (Zugriff am 1. Oktober 2022).

[17] L. Wei, B. Luo, Y. Li, Y. Liu und Q. Xu, „I Know What You See“, in *ACSAC '18: 2018 Annual Computer Security Applications Conference*, San Juan PR USA. New York, NY, USA: ACM, 2018. Zugriff am: 4. Oktober 2022. [Online]. Verfügbar unter: https://doi.org/10.1145/3274694.3274696