Cybersecurity and Law

2025 Nr 2 (14)

DOI: 10.34567/cal/215566

Taxonomy of Contemporary Attacks on Machine Learning Models

Taksonomia Współczesnych Ataków na Modele Uczenia Maszynowego

Kacper ZDROJEWSKI
Akademia Sztuki Wojennej
ORCID: 0009-0007-5255-0441
E-mail: k.zdrojewski@doktorant.akademia.mil.pl

Abstract

This paper analyzes security threats to artificial intelligence (AI) systems, focusing on the classification of technical attacks that target confidentiality, integrity, and availability of machine learning (ML) models.

The study is based on a systematic review of scientific literature and an examination of established taxonomies, including those by NIST, ENISA, and MITRE ATLAS, enabling the integration of attack types across stages of the ML lifecycle and attacker knowledge levels.

The work identifies major threat vectors such as adversarial examples, data poisoning, backdoor insertion, model theft, membership inference, model overloading, and prompt manipulation. The results show that white-box attacks allow highly precise and covert manipulations of model behavior, while black-box attacks dominate production environments, leveraging transferability and prediction-based analysis.

The proposed taxonomy demonstrates that securing ML systems requires a multilayered defensive approach, encompassing protection of training data, strengthening model robustness, and securing API interfaces using techniques such as differential privacy, adversarial training, and rate limiting. The findings highlight the need for close collaboration between machine learning engineers and cybersecurity practitioners to ensure resilient and trustworthy AI systems.

Keywords
artificial intelligence; adversarial attacks; threat taxonomy; white-box; black-box

Streszczenie

Niniejsza praca analizuje zagrożenia bezpieczeństwa systemów sztucznej inteligencji (AI), koncentrując się na klasyfikacji technicznych ataków ukierunkowanych na naruszenie poufności, integralności i dostępności modeli uczenia maszynowego (ML).

Oparto ją na systematycznym przeglądzie publikacji naukowych oraz analizie istniejących ram klasyfikacyjnych, w tym NIST, ENISA i MITRE ATLAS, co umożliwiło zestawienie dominujących typów ataków w kontekście etapów cyklu życia systemu ML oraz poziomu wiedzy przeciwnika.

Praca identyfikuje kluczowe wektory zagrożeń, takie jak ataki adwersarialne, zatruwanie danych, backdoory, kradzież modelu, inferencja członkostwa, przeciążanie modeli oraz manipulacja promptami. Wyniki wskazują, że ataki white-box umożliwiają precyzyjne, trudne do wykrycia modyfikacje modelu, natomiast ataki black-box dominują w środowiskach produkcyjnych i bazują na zjawisku przenaszalności oraz analizie predykcji.

Przedstawiona taksonomia prowadzi do wniosku, że bezpieczeństwo systemów ML wymaga podejścia warstwowego, obejmującego ochronę danych treningowych, wzmacnianie odporności modeli oraz zabezpieczanie interfejsów API za pomocą technik takich jak differential privacy, adversarial training i rate limiting. Rezultaty podkreślają konieczność ścisłej współpracy zespołów ML i cyberbezpieczeństwa w celu budowy odpornych i godnych zaufania systemów AI.

Słowa kluczowe

sztuczna inteligencja; ataki adwersarialne; taksonomia zagrożeń; white-box; black-box

Wprowadzenie

Współczesna technologia charakteryzuje się gwałtownym rozwojem i masową komercjalizacją systemów sztucznej inteligencji (AI), które stały się integralną częścią krytycznych sektorów gospodarki. W centrum tego zjawiska znajdują się metody uczenia maszynowego (ML) oraz głębokiego uczenia (DL). Uczenie maszynowe to poddziedzina AI, która umożliwia komputerom uczenie się z danych i identyfikację wzorców, bez konieczności jawnego programowania reguł działania. Zamiast instrukcji, algorytm otrzymuje duży zbiór danych, na podstawie którego tworzy model zdolny do podejmowania decyzji lub przewidywań (np. klasyfikacja wiadomości jako spam). Głębokie uczenie jest natomiast specjalistyczną formą ML, która wykorzystuje głębokie sieci neuronowe (architektury z wieloma warstwami), aby przetwarzać dane w bardziej abstrakcyjny sposób i z większą skutecznością, szczególnie w zadaniach związanych z obrazem, mową i językiem naturalnym. Ta wszechobecność systemów ML i DL generuje nowe, unikalne wyzwania dla cyberbezpieczeństwa, wykraczające poza tradycyjne modele ochrony danych i systemów informatycznych. Bezpieczeństwo AI, często określane jako AISec lub MLSec, koncentruje się na zapewnieniu poufności, integralności i dostępności samego modelu oraz danych, na których został wytrenowany[1]. W przeciwieństwie do tradycyjnego oprogramowania, modele ML są podatne na ataki, które wykorzystują ich statystyczny charakter i wrażliwość na subtelne manipulacje danych wejściowych.

Celem niniejszej pracy jest stworzenie szczegółowej i ustrukturyzowanej taksonomii technicznych ataków skierowanych przeciwko modelom uczenia maszynowego oraz dogłębna charakterystyka ich mechanizmów działania, ze wskazaniem kluczowych wyzwań dla dziedziny cyberbezpieczeństwa. Głównym problemem badawczym jest pytanie: Jakie są główne wektory i techniki wykorzystywane do naruszenia poufności, integralności i dostępności modeli ML

[1] K. Liderman, Bezpieczeństwo informacyjne, Warszawa 2017.

oraz w jaki sposób należy je sklasyfikować pod względem metodyki i celu ich ataku.

W pracy zastosowano metodę systematycznego przeglądu literatury oraz analizy treści. Zostały przeanalizowane kluczowe publikacje naukowe z wiodących konferencji i czasopism z zakresu bezpieczeństwa i uczenia maszynowego. Zebrane techniki ataków zostały poddane syntezie i klasyfikacji w celu stworzenia taksonomii opartej na celach ataku i naruszanych zasadach bezpieczeństwa, co zapewnia kompleksowe i ustrukturyzowane ujęcie problemu dla odbiorców o różnym poziomie wiedzy technicznej.

Przegląd literatury

Wraz z wykładniczym wzrostem znaczenia systemów uczenia maszynowego w krytycznych zastosowaniach, społeczność naukowa i sektor cyberbezpieczeństwa intensywnie pracują nad systematyzacją i klasyfikacją związanych z nimi zagrożeń. Istniejące ramy taksonomiczne różnią się kryteriami, skupiając się bądź to na fazie cyklu życia modelu, celu ataku, czy też na wiedzy i możliwościach technicznych atakującego.

Jedną z najbardziej szczegółowych i praktycznych perspektyw w kontekście taktyk i technik cyberprzestępczych jest macierz MITRE ATLAS[2]. Macierz ta stanowi bazę wiedzy o taktykach, technikach i procedurach stosowanych przez przeciwników w atakach na systemy AI. Podobnie, jak w przypadku macierzy ATT&CK dla tradycyjnych systemów informatycznych, ATLAS opisuje realne przypadki użycia, w których systemy uczenia maszynowego są albo celem, albo narzędziem ataku. Obejmuje ona około 14 taktyk i ponad 50 różnych technik, takich jak zatruwanie danych (data poisoning), omijanie modelu (model evasion) i kompromitacja łańcucha dostaw[3].

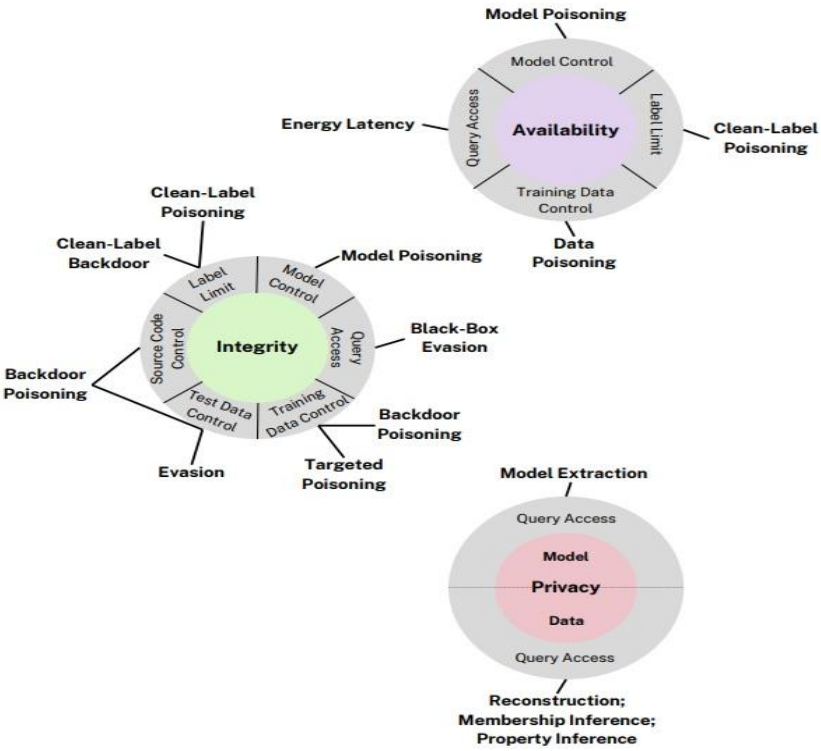
Z kolei National Institute of Standards and Technology (NIST) w raporcie AI 100-2 E2025[4] proponuje taksonomię opartą na hierarchii pojęć, która obejmuje rodzaje metod ML, etapy cyklu życia ataku oraz cele, możliwości i wiedzę atakującego. NIST kładzie szczególny nacisk na ujednolicenie terminologii, klasyfikując ataki na Predictive AI (Rys. 1) i Generative AI (Rys. 2), w tym ataki omijające, zatruwające i naruszające prywatność. W kontekście ataków omijających raport wyróżnia podejścia black-box (atakujący ma dostęp tylko do wyników predykcji) oraz white-box (znana jest architektura i wagi modelu).

[2] MITRE Corporation, ATLAS: Adversarial Threat Landscape for Artificial-Intelligence Systems, https://atlas.mitre.org, [dostęp: 28.11.2025].

[3] Vide: MITRE Corporation, ATLAS Matrix, https://atlas.mitre.org/matrices/ATLAS, [dostęp: 28.11.2025].

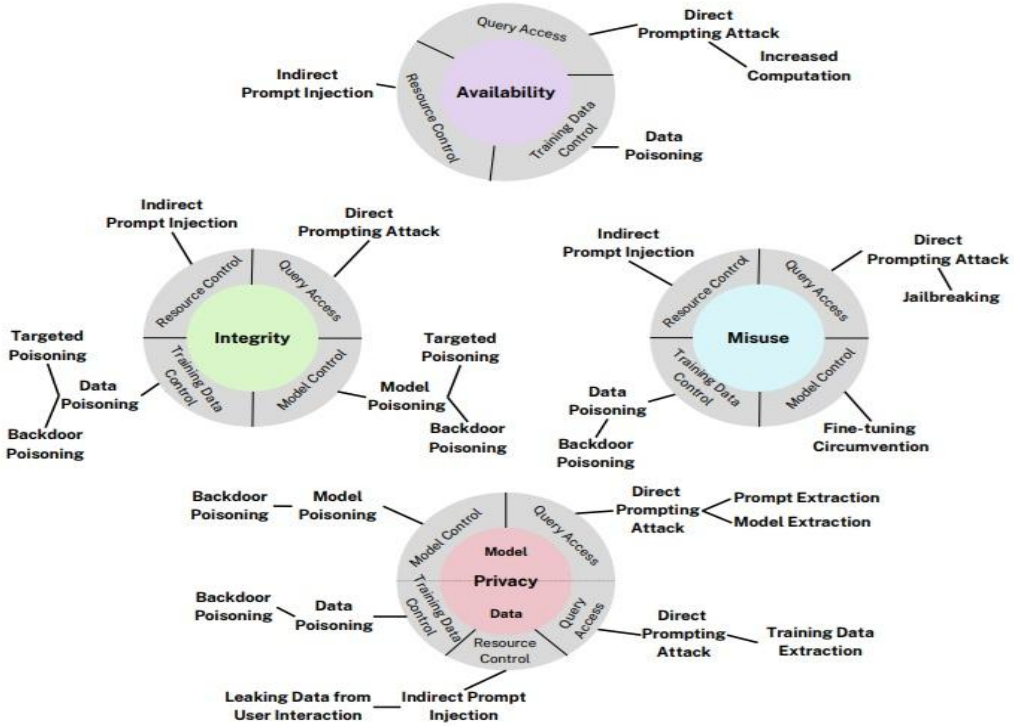
[4] A. Vassilev, A. Oprea, A. Fordyce, H. Anderson, X. Davies, M. Hamin, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, (National Institute of Standards and Technology, Gaithersburg, MD) NIST Trustworthy and Responsible AI 2025, NIST AI 100-2e2025, https://doi.org/10.6028/NIST.AI.100-2e2025.

Rys. 1. *Taksonomia ataków na predykcyjne systemy AI*



Źródło: *Raport NIST AI 100-2 E2025, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*.

Rys. 2. *Taksonomia ataków na generatywne systemy AI*



Źródło: *Raport NIST AI 100-2 E2025, Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations*.

Raport ENISA Securing Machine Learning Algorithms[5] wnosi do dyskusji o atakach na modele AI uporządkowaną i całościową perspektywę bezpieczeństwa uczenia maszynowego, obejmującą zarówno klasyfikację zagrożeń oraz ich mapowanie na etapy cyklu życia modelu. Dokument identyfikuje kluczowe wektory ataku - w szczególności zatruwanie danych treningowych, manipulacje adwersarialne, wycieki informacji oraz ingerencje w proces uczenia - pokazując, że podatności pojawiają się na każdym etapie, od pozyskania danych po wdrożenie i eksploatację systemu. ENISA podkreśla również konieczność stosowania wielowarstwowych środków ochronnych oraz doboru mechanizmów bezpieczeństwa niewpływających istotnie na jakość działania modeli.

Kluczowe zagrożenia dotyczące bezpieczeństwa modeli uczenia maszynowego oraz ich podatności na ataki na różnych etapach funkcjonowania, wraz z pokazem praktycznym, zostały przedstawione w prezentacji pt. „Zagrożenia dla modeli uczenia maszynowego: typologia ataków[6]" zaprezentowanej podczas konferencji CyberEXPERT 2024. Uzupełnieniem tej prezentacji jest kompleksowa tabela prezentująca typy ataków na sztuczną inteligencję wraz z ich opisem, konsekwencjami, a także metodami obrony[7].

Dalszą techniczną głębię zapewnia artykuł „Depth, Breadth, and Complexity: Ways to Attack and Defend Deep Learning Models[8]" oraz inne publikacje udostępnione przez AI and Secure Computing Research Lab Uniwersytetu Chicagowskiego[9]. Prace te szczegółowo analizują, w jaki sposób złożoność modeli głębokiego uczenia wpływa na ich podatności, oferując jednocześnie pogłębione omówienie mechanizmów generowania przykładów adwersarialnych oraz technik obronnych w scenariuszach black-box i white-box. Przegląd literatury pozwala zatem na przyjęcie metodycznego podejścia łączącego klasyfikację opartą na celach ataku w cyklu życia AI (dane, algorytm, wyniki) z kryterium zakresu wiedzy atakującego (black-box vs. white-box). Taka dwuwymiarowa taksonomia umożliwia precyzyjniejszą charakterystykę mechanizmów obronnych oraz dokładniejsze definiowanie wektorów zagrożeń.

Architektura systemu uczenia maszynowego

Przed przejściem do szczegółowej taksonomii ataków, konieczne jest zdefiniowanie kluczowych elementów cyklu życia typowego systemu uczenia

[5] European Union Agency for Cybersecurity (ENISA), Securing machine learning algorithms, (2021, December 14), https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20-%20Securing%20Machine%20Learning%20Algorithms.pdf, [dostęp: 28.11.2025].
[6] M. Malinowski, Zagrożenia dla modeli uczenia maszynowego: typologia ataków, CyberEXPERT 24, Eksperckie Centrum Szkolenia Cyberbezpieczeństwa, 5-6 Listopada 2024 r., https://doi.org/10.6084/m9.figshare.27908265.v1.
[7] M. Malinowski, Typy ataków na Sztuczną Inteligencję, https://spaces-cdn.owlstown.com/blobs/t60xgrk64ey2uqq3qe7oe6kn5aux, [dostęp: 28.11.2025].
[8] F. Juraev, E. Abdukhamidov, M. Abuhamad, T. Abuhmed, Depth, Breadth, and Complexity: Ways to Attack and Defend Deep Learning Models, In Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security (ASIA CCS'22), Association for Computing Machinery, New York, USA, pp. 1207-1209, https://doi.org/10.1145/3488932.3527278.
[9] AI and Secure Computing Research Lab, Loyola University Chicago, https://aisec.cs.luc.edu, [dostęp: 28.11.2025].

maszynowego. Każdy z nich stanowi potencjalny wektor ataku, co tworzy fundament dalszej klasyfikacji zagrożeń.

Dane treningowe to surowe dane (np. obrazy, teksty, dane finansowe), które służą algorytmowi do nauki i budowania wzorców. Jakość, reprezentatywność i poprawność danych treningowych decydują o zdolności modelu do podejmowania poprawnych decyzji w przyszłości. Stanowią one pierwszy i najbardziej podstawowy wektor ataku. Jeżeli dane zostaną celowo zmanipulowane (np. poprzez dodanie złośliwych próbek), może to prowadzić do zatrucia modelu. Ponadto, ponieważ dane te często zawierają poufne informacje, są one celem ataków nastawionych na naruszenie prywatności.

Algorytm uczenia to zestaw metod optymalizacyjnych (np. spadku gradientu) oraz reguł matematycznych, które przetwarzają dane treningowe w celu wyznaczenia parametrów modelu (funkcji decyzyjnej). Algorytm ten jest kluczowy w fazie trenowania, ponieważ odpowiada za efektywną naukę. Choć ataki bezpośrednio na metody optymalizacyjne są rzadsze, błędna konfiguracja hiperparametrów lub manipulacja środowiskiem treningowym może prowadzić do powstania modeli o obniżonej odporności na ataki adwersarialne lub ukrytych podatnościach.

Model jest ostatecznym rezultatem pracy algorytmu na danych treningowych – można go traktować jako „wytrenowany mózg” systemu AI. To zbiór parametrów (wag i biasów) pozwalający na wykonywanie przewidywań na nowych danych. Model jest własnością intelektualną twórców i często zawiera skompresowaną wiedzę pochodzącą ze zbioru treningowego. Z tego względu jest głównym celem ataków ekstrakcji, w których atakujący próbuje stworzyć jego funkcjonalną kopię.

Dane wejściowe to nowe przypadki, które trafiają do modelu już po zakończeniu procesu trenowania i wdrożeniu go do środowiska produkcyjnego. Te dane nie były uwzględnione podczas fazy uczenia. Model przetwarza dane wejściowe, wykorzystując wcześniej zidentyfikowane zależności, co pozwala na ocenę zdolności modelu do generalizowania wyuczonych wzorców. Dane wejściowe są głównym celem ataków omijających, gdzie celowo spreparowane dane mają za zadanie wymusić na modelu niepoprawną predykcję, naruszając integralność jego działania.

Wyniki lub predykcje to decyzje, klasy lub wartości generowane przez model po przetworzeniu danych wejściowych. Wynik jest jedynym elementem systemu, do którego atakujący ma bezpośredni i łatwy dostęp w przypadku modelu typu black-box. Poprzez analizę tych odpowiedzi, a zwłaszcza stopnia pewności (confidence) modelu co do swojej predykcji, atakujący może wyciągać wnioski o wewnętrznej architekturze modelu lub nawet o tym, jakie konkretne rekordy danych zostały użyte do jego trenowania.

Zrozumienie tych elementów umożliwia precyzyjne mapowanie wektorów zagrożeń na poszczególne fazy cyklu życia systemów ML - od pozyskiwania i przygotowania danych, przez proces trenowania, aż po etap wdrożenia i generowania predykcji.

Ataki typu white-box

Ataki w scenariuszu white-box stanowią najbardziej zaawansowane zagrożenie dla modeli uczenia maszynowego, ponieważ przeciwnik posiada pełną wiedzę o wewnętrznej architekturze i mechanice systemu. Dostęp ten

obejmuje kod źródłowy, architekturę modelu (np. typ sieci neuronowej, liczba warstw), wagi i parametry (zbiór wszystkich liczb, które model wyuczył), funkcję straty oraz często zbiór danych treningowych. Taki poziom dostępu pozwala atakującemu na wykorzystanie technik opartych na gradientach i optymalizacji matematycznej, prowadząc do ataków o wysokiej skuteczności.

Najbardziej powszechnym wektorem ataku white-box, skierowanym na integralność modelu w fazie wnioskowania, jest generowanie adwersarialnych przykładów omijających (Evasion Attacks). Unikalność tej metody polega na tym, że atakujący, mając pełną wiedzę o nachyleniu powierzchni decyzyjnej modelu (czyli jego gradientach), może obliczyć minimalną, niezauważalną dla ludzkiego oka perturbację (szum), którą należy dodać do danych wejściowych, aby wymusić błędną, z góry określoną predykcję. W konsekwencji można skutecznie oszukać systemy krytyczne (np. autonomiczny pojazd błędnie odczytujący znak drogowy) przy zachowaniu pozornej nieszkodliwości danych. Stopień trudności technicznej jest niski do umiarkowanego, ale wymaga dostępu do wewnętrznych zasobów systemu. Kluczowe metody w tej grupie, które zrewolucjonizowały dziedzinę, to L-BFGS[10] (jedna z pierwszych metod opartych na optymalizacji), PGD[11] (Projected Gradient Descent), uznawany za złoty standard w testowaniu odporności modeli, oraz szybka, jednokrokowa technika FGSM[12] (Fast Gradient Sign Method). Ataki te mogą być ukierunkowane – wymuszając konkretną, błędną klasę, lub nieukierunkowane – wymuszając dowolną inną, niepoprawną klasę.

W ramach ataków omijających możliwe są również te, które realizuje się w świecie fizycznym. Ich charakterystyczną cechą jest projektowanie perturbacji w sposób zapewniający odporność na zniekształcenia środowiskowe, takie jak zmiana kąta widzenia, oświetlenia czy perspektywy kamery. Przykładowo, atak może polegać na wydrukowaniu oprawek okularów wprowadzających w błąd system rozpoznawania twarzy[13] albo na umieszczeniu czarno-białych naklejek na znakach drogowych w celu oszukania klasyfikatora pojazdu autonomicznego[14]. Konsekwencją jest przeniesienie zagrożenia z domeny cyfrowej do fizycznej, co generuje bezpośrednie ryzyko dla bezpieczeństwa. Ataki tego typu są trudniejsze do przeprowadzenia, gdyż wymagają uwzględnienia dodatkowych ograniczeń związanych z fizyczną implementacją i trwałością perturbacji.

Ataki typu backdoor (Backdoor Attacks) to jeden z najbardziej podstępnych ataków white-box na integralność modelu. Unikalność tej metody

[10] D.C. Liu, J. Nocedal, On the limited memory BFGS method for large scale optimization, “Mathematical Programming 45” 1989, s. 503–528, https://doi.org/10.1007/BF01589116.

[11] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, A. Vladu, Towards deep learning models resistant to adversarial attacks, 6th International Conference on Learning Representations, ICLR 2018, Vancouver, Canada, April 30 - May 3 2018, https://doi.org/10.48550/arXiv.1706.06083.

[12] I. Goodfellow, J. Shlens, C. Szegedy, Explaining and harnessing adversarial examples, International Conference on Learning Representations, 2015, https://doi.org/10.48550/arXiv.1412.6572.

[13] M. Sharif, S. Bhagavatula, L. Bauer, M.K. Reiter, Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security, October 2016, https://doi.org/10.1145/2976749.2978392.

[14] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, D. Song, Robust physical-world attacks on deep learning visual classification, 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, USA, June 18-22, 2018, s. 1625–1634, https://doi.org/10.1109/CVPR.2018.00175.

polega na umieszczeniu ukrytej funkcji w modelu, która pozostaje nieaktywna podczas normalnego działania. Zostaje ona aktywowana wyłącznie przez specyficzny, ukryty wzorzec (trigger) wprowadzony przez atakującego. Atakujący, mając pełną wiedzę o algorytmie i danych treningowych, może precyzyjnie kontrolować ten proces. W konsekwencji udanego ataku, przeciwnik może całkowicie przejąć kontrolę nad systemem lub sabotować jego działanie, wprowadzając błędy tylko w ściśle określonych warunkach. Stopień trudności takiego ataku jest wysoki, ponieważ wymaga dostępu do modelu i jego modyfikacji przed wdrożeniem.

Zatrucie danych treningowych (Data Poisoning) to klasyczny atak white-box na integralność i pierwotną fazę cyklu życia ML. Wyróżnia go to, że złośliwe dane są wprowadzane do zbioru treningowego, co trwale uszkadza proces trenowania i prowadzi do błędnego działania całego modelu. Pełna wiedza o danych i algorytmie pozwala na takie manipulacje, które mogą celowo wypaczyć wyniki modelu (Model Skewing). W konsekwencji model zwraca stronnicze lub nieprawidłowe predykcje, co prowadzi do błędnych decyzji operacyjnych. Stopień trudności jest niski do umiarkowanego, lecz wymaga kontroli nad źródłami danych lub możliwością wstrzyknięcia złośliwych próbek do zbioru treningowego.

Luki w łańcuchu dostaw (Supply Chain Vulnerabilities) to ogólny atak, który często przyjmuje formę white-box w momencie kompromitacji wewnętrznych komponentów lub danych dostarczanych do systemu AI. Cechą wyróżniającą jest to, że atak ten wykorzystuje słabości w komponentach (np. bazowych modelach, bibliotekach), prowadząc do naruszenia integralności i poufności. W konsekwencji może skutkować awarią systemu, wyciekiem danych i zakłóceniami operacyjnymi, przenosząc ryzyko z zewnętrznego środowiska do wewnętrznych operacji. Stopień trudności jest umiarkowany, zależy od poziomu zabezpieczeń procesów kontroli dostaw.

Ataki typu black-box

Ataki w scenariuszu black-box zakładają, że przeciwnik ma ograniczony dostęp do modelu z możliwością jedynie interakcji z nim za pomocą publicznie dostępnego interfejsu API (wysyłanie zapytań i odbieranie predykcji). Atakujący nie zna architektury, wag ani zbioru danych treningowych. Ataki te opierają się na ekstrakcji informacji i zjawisku przenaszalności luk.

Ataki omijające w scenariuszu black-box są atakami na integralność, gdzie wykorzystuje się zmodyfikowane dane wejściowe w celu uniknięcia wykrycia przez system AI. Unikalność tej metody polega na wykorzystaniu zjawiska przenaszalności (transferability): złośliwy przykład wygenerowany na lokalnym modelu zastępczym jest z sukcesem aplikowany do docelowego modelu black-box. W konsekwencji atakujący omija systemy wykrywania (np. filtry spamowe, systemy detekcji intruzów), co może skutkować przepuszczeniem zagrożeń bez odpowiedniej reakcji.

Kradzież modelu (Model Theft) to kluczowy atak black-box na poufność modelu jako własności intelektualnej. Cechą wyróżniającą jest to, że atakujący systematycznie wysyła dużą liczbę zapytań do API modelu docelowego. Na podstawie otrzymywanych wyników trenuje lokalny model zastępczy (substitute

model[15]), który funkcjonalnie naśladuje oryginał. W konsekwencji następuje kradzież własności intelektualnej, utrata przewagi konkurencyjnej, a skradziony klon może zostać wykorzystany do przeprowadzenia łatwiejszych, wewnętrznych ataków white-box.

Atak inferencji członkostwa (Membership Inference Attacks) to atak black-box na poufność danych treningowych. Wyróżnia go to, że polega na analizie stopnia pewności predykcji modelu (bez dostępu do jego wnętrza) w celu ustalenia, czy konkretny rekord danych został użyty do jego trenowania. Atakujący trenuje model cieniowy (shadow model[16]), który uczy się, kiedy model docelowy jest „nadmiernie pewny", co jest wskaźnikiem na to, że widział dany rekord. W konsekwencji dochodzi do ujawnienia informacji o danych, co może naruszyć poufność i prywatność (np. w kontekście danych medycznych). Stopień trudności jest umiarkowany i wymaga zasobów do wytrenowania modelu cieniowego.

Wstrzyknięcie polecenia (Prompt Injection) jest krytycznym atakiem black-box na integralność, szczególnie istotnym dla dużych modeli językowych (LLM). Cechą wyróżniającą jest to, że atakujący manipuluje danymi wejściowymi (komendami/promptami), aby uzyskać nieautoryzowany dostęp lub kontrolę nad modelem. W konsekwencji model wykonuje nieautoryzowane działania (np. ujawnia poufne dane, generuje szkodliwą treść, ignoruje systemowe instrukcje bezpieczeństwa). Stopień trudności jest niski i często sprowadza się do umiejętnej inżynierii promptów, bez potrzeby zaawansowanej wiedzy o ML.

Atak typu odmowa usługi (Model Denial of Service) jest atakiem black-box na dostępność modelu. Wyróżnia go celowe przeciążenie modelu operacjami wymagającymi dużych zasobów (np. długie, złożone zapytania do LLM lub wielokrotne zapytania do modelu analitycznego). W konsekwencji dochodzi do zablokowania dostępu do zasobów modelu, co prowadzi do przerw w jego działaniu, utraty dostępności dla legalnych użytkowników i wzrostu kosztów operacyjnych. Stopień trudności jest niski i może wymagać jedynie dużej liczby prostych zapytań.

Synteza taksonomii: mapowanie ataków na cykl życia systemu ML

Dotychczasowa analiza pozwoliła wyodrębnić ataki na modele uczenia maszynowego, według dwóch kluczowych kryteriów: poziomu wiedzy i dostępu atakującego oraz elementów architektury systemu. Niniejszy rozdział ma na celu syntezę tych klasyfikacji, prezentując kompleksowe mapowanie, które ujawnia, w jaki sposób różne typy zagrożeń są ukierunkowane na konkretne fazy życia systemu ML. Zrozumienie tego dwuwymiarowego podejścia jest kluczowe dla

[15] Główna rola modelu zastępczego to: odtworzyć funkcjonalność modelu docelowego (jego decyzje, granice decyzyjne), nie zaś jego błędy generalizacji czy poziom przetrenowania. Vide: N. Papernot, P. McDaniel, I. Goodfellow, S. Jha, Z. B. Celik, A. Swami, Practical black-box attacks against machine learning, Proceedings of the 2017 ACM on Asia conference on computer and communications security, 2017, s. 506-519, https://doi.org/10.48550/arXiv.1602.02697.

[16] Trenowany po to, aby odtworzyć sposób, w jaki model docelowy generalizuje i przetrenowuje się (różnice w pewności predykcji dla danych widzianych i niewidzianych). Vide: R. Shokri, M. Stronati, C. Song, V. Shmatikov, Membership inference attacks against machine learning models, 2017 IEEE symposium on security and privacy, s. 3-18, https://doi.org/10.48550/arXiv.1610.05820.

projektowania efektywnych strategii obronnych, które muszą być wdrożone w każdym punkcie krytycznym. Tabela 1 porządkuje kluczowe wektory ataku opisane w niniejszej pracy.

Tabela 1. *Mapowanie kluczowych ataków na systemy ML według fazy cyklu życia i wiedzy atakującego*

Typ ataku	Wektor ataku	Cel ataku	Wiedza atakującego
Ataki Adwersarialne/Omijające (np. PGD, FGSM)	Dane Wejściowe	Integralność (precyzyjne oszukanie predykcji)	white-box
Ataki typu backdoor	Model	Integralność (przejęcie kontroli w określonych warunkach)	white-box
Zatrucie danych treningowych	Dane Treningowe	Integralność (trwałe błędy operacyjne/stronniczość)	white-box
Luki w łańcuchu dostaw	Dane Treningowe/Model	Integralność i Poufność (kompromitacja komponentów)	white-box
Ataki Omijające (Przenaszalność/Query)	Dane Wejściowe	Integralność (ominięcie detekcji)	black-box
Kradzież modelu	Model	Poufność (kradzież własności intelektualnej)	black-box
Atak inferencji członkostwa	Dane Treningowe	Poufność (naruszenie prywatności danych)	black-box
Wstrzyknięcie polecenia	Dane Wejściowe	Integralność (manipulacja modelem, np. LLM)	black-box
Atak typu odmowa usługi	Model	Dostępność (przerwanie działania)	black-box

Źródło: *opracowanie własne*.

Należy podkreślić, że niniejsza praca koncentruje się na wybranych, najważniejszych i najczęściej występujących wektorach ataku na systemy uczenia maszynowego, które zostały zidentyfikowane jako kluczowe zagrożenia w obecnym krajobrazie cyberbezpieczeństwa AI. Obszar ten charakteryzuje się jednak dynamicznym rozwojem, a liczba zidentyfikowanych taktyk i technik jest znacznie szersza, obejmując również zaawansowane metody, takie jak ataki na transfer uczenia (Transfer Learning Attacks) czy ataki bocznokanałowe (Side-Channel Attacks)[17], lecz wykracza to poza ramy niniejszego studium.

Konkluzje

[17] M. Malinowski, Typy ataków na Sztuczną Inteligencję, https://spaces-cdn.owlstown.com/blobs/t60xgrk64ey2uqq3qe7oe6kn5aux, [dostęp: 28.11.2025].

Niniejsza praca przedstawiła taksonomię zagrożeń dla systemów uczenia maszynowego, opartą na dwóch kluczowych kryteriach: poziomie wiedzy atakującego (white-box i black-box) oraz fazie cyklu życia systemu ML, która stanowi cel ataku (dane treningowe, model, dane wejściowe, wyniki). Przeprowadzona analiza wykazała, że bezpieczeństwo systemów ML jest wyzwaniem wielowymiarowym, wymagającym holistycznego podejścia, które obejmuje każdy etap projektowania, wdrażania i eksploatacji modelu.

Przedstawiona synteza potwierdza, że zagrożenia dla systemów AI są wszechobecne, ale ich charakter zależy od dostępu, jakim dysponuje przeciwnik. Ataki typu white-box charakteryzują się maksymalną precyzją i trwałością wprowadzanych uszkodzeń. Pełna wiedza o architekturze modelu umożliwia atakującym optymalizację celów złośliwego działania, zagrażając integralności i poufności w sposób trudny do wykrycia zewnętrznymi metodami. Z kolei ataki typu black-box stanowią najczęstsze zagrożenie w środowiskach produkcyjnych, gdzie model jest udostępniany jako usługa API. Ich skuteczność opiera się na eksploatacji zewnętrznych interfejsów oraz zjawiskach takich jak przenaszalność luk, umożliwiając naruszenie poufności i dostępności bez bezpośredniego wglądu w model. Kluczową obserwacją jest to, że pomimo braku wiedzy, ataki black-box stanowią realne zagrożenie, wymuszając na obrońcach stosowanie technik monitorowania ruchu sieciowego i walidacji zapytań zamiast polegania na wewnętrznych zabezpieczeniach modelu.

Główną implikacją płynącą z taksonomii jest konieczność wdrożenia obrony warstwowej, która odpowiada specyfice zagrożeń w poszczególnych fazach cyklu życia ML:

- zabezpieczenie danych treningowych: w fazie treningu kluczowe jest stosowanie technik prywatności różnicowej (Differential Privacy[18]) oraz ścisłe audytowanie pochodzenia danych w celu ochrony przed zatruciem danych i atakami inferencji członkostwa;
- wzmacnianie odporności modelu: model musi być testowany pod kątem odporności adwersarialnej za pomocą technik takich jak Adversarial Training[19], aby skutecznie przeciwdziałać atakom adwersarialnym i atakom typu backdoor;
- ochrona interfejsu API: jako główna brama dla większości ataków black-box, musi być zabezpieczony za pomocą rate limiting[20] (przeciwdziałanie DoS) i zaawansowanych mechanizmów walidacji wejść.

Podsumowując, zapewnienie bezpieczeństwa systemów ML nie jest jednorazowym zadaniem, lecz ciągłym procesem wymagającym współpracy między inżynierami uczenia maszynowego, ekspertami ds. cyberbezpieczeństwa i legislatorami, aby stworzyć odporne, godne zaufania systemy AI.

Bibliografia

[18] Mechanizm kryptograficzny dodający kontrolowany szum do danych lub wyników modelu, tak aby zminimalizować ryzyko identyfikacji pojedynczego rekordu danych.

[19] Polega na celowym wprowadzaniu adwersarialnych przykładów (próbek z dodanym złośliwym szumem) do zbioru treningowego i ponownym trenowaniu modelu, ucząc go tym samym, jak poprawnie klasyfikować zarówno oryginalne, jak i zniekształcone dane.

[20] Technika polegająca na ograniczaniu liczby zapytań (requestów) do modelu w określonym czasie.

AI and Secure Computing Research Lab, Loyola University Chicago, https://aisec.cs.luc.edu.

European Union Agency for Cybersecurity (ENISA), Securing machine learning algorithms, (2021, December 14), https://www.enisa.europa.eu/sites/default/files/publications/ENISA%20Report%20%20Securing%20Machine%20Learning%20Algorithms.pdf.

Eykholt K., Evtimov I., Fernandes E., Li B., Rahmati A., Xiao C., Prakash A., Kohno T., Song D., Robust physical-world attacks on deep learning visual classification, 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, USA, June 18-22, 2018, https://doi.org/10.1109/CVPR.2018.00175.

Goodfellow I., Shlens J., Szegedy C., Explaining and harnessing adversarial examples, International Conference on Learning Representations, 2015, https://doi.org/10.48550/arXiv.1412.6572.

Juraev F., Abdukhamidov E., Abuhamad M., Abuhmed T., Depth, Breadth, and Complexity: Ways to Attack and Defend Deep Learning Models, In Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security (ASIA CCS '22), Association for Computing Machinery, New York, USA, https://doi.org/10.1145/3488932.3527278.

Liderman K., Bezpieczeństwo informacyjne, Warszawa 2017.

Liu D.C., Nocedal J., On the limited memory BFGS method for large scale optimization, “Mathematical Programming 45” 1989, https://doi.org/10.1007/BF01589116.

Madry A., Makelov A., Schmidt L., Tsipras D., Vladu A., Towards deep learning models resistant to adversarial attacks, 6th International Conference on Learning Representations, ICLR 2018, Vancouver, Canada, April 30 - May 3, 2018, https://doi.org/10.48550/arXiv.1706.06083.

Malinowski M., Typy ataków na Sztuczną Inteligencję, https://spaces-cdn.owlstown.com/blobs/t60xgrk64ey2uqq3qe7oe6kn5aux.

Malinowski M., Zagrożenia dla modeli uczenia maszynowego: typologia ataków, CyberEXPERT 24, Eksperckie Centrum Szkolenia Cyberbezpieczeństwa, 5-6 Listopada 2024 r., https://doi.org/10.6084/m9.figshare.27908265.v1.

MITRE Corporation, ATLAS: Adversarial Threat Landscape for Artificial Intelligence Systems, https://atlas.mitre.org/matrices/ATLAS.

Papernot N., McDaniel P., Goodfellow I., Jha S., Celik Z. B., Swami A., Practical black-box attacks against machine learning, Proceedings of the 2017 ACM on Asia conference on computer and communications security, 2017, https://doi.org/10.48550/arXiv.1602.02697.

Pilarski G., Przegląd współczesnych zagrożeń w cyberprzestrzeni, „Przegląd Techniki Uzbrojenia” 2023, t. 167(5), https://doi.org/10.5604/01.3001.0054.2998.

Sharif M., Bhagavatula S., Bauer L., Reiter M.K., Accessorize to a crime: Real and stealthy attacks on state-of-the-art face recognition, Proceedings of the 23rd ACM SIGSAC Conference on Computer and Communications Security, October 2016, https://doi.org/10.1145/2976749.2978392.

Shokri R., Stronati M., Song C., Shmatikov V., Membership inference attacks against machine learning models, 2017 IEEE symposium on security and privacy, https://doi.org/10.48550/arXiv.1610.05820.

Vassilev A., Oprea A., Fordyce A., Anderson H., Davies X., Hamin M., Adversarial Machine Learning: A Taxonomy and Terminology of Attacks and Mitigations, (National Institute of Standards and Technology, Gaithersburg, MD) NIST Trustworthy and Responsible AI 2025, NIST AI 100-2e2025, https://doi.org/10.6028/NIST.AI.100-2e2025.