



**Vilniaus
universitetas**

Mašininiu mokymusi grindžiami metodai apgaulingoms ir obfuskuotoms kenkėjiškoms programoms generuoti stiprinant kibernetinį saugumą

Doktorantūros pradžios/pabaigos metai: 2023–2027

Studijų metai: 2025–2026 antras pusmetis

Mokslo kryptis: Informatikos inžinerija T 007

Doktorantas: Juozas Dautartas

Vadovas: Dr. Viktor Medvedev

Studijų planas ir jo vykdymo suvestinė

Studijų metai	Egzaminai ¹	
	Planas	Įvykdyta
I (2023/2024)	3	3
II (2024/2025)	1	1
III (2025/2026)		
IV (2026/2027)		
Iš viso:	4	4

Studijų metai	Dalyvavimas konferencijose				Publikacijos					
	Tarptautinėse ²		Nacionalinėse ³		Su citav. rodikliu ⁴			Be citav. rodiklio ⁵		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta ⁶	Būklė ⁷	Planas	Įvykdyta ⁶	Būklė ⁷
I (2023/2024)				1						
II (2024/2025)			1	1						
III (2025/2026)	1	1		1	1	0	Įteikta	1	2	Priimta
IV (2026/2027)	1				1					
Iš viso:	2	1	1	3	2	0		1	2	

Ataskaitinio pusmečio darbo planas ir jo vykdymo suvestinė

Vilniaus
universitetas

Egzaminai 2025/2026 (II pusmetis)		
Planas	Ivykdyta	Būklė
-	-	-

Dalyvavimas konferencijose 2025/2026 (II pusmetis)		
Planas	Ivykdyta	Konferencijos tipas
-	-	-

Publikacijos 2025/2026 (II pusmetis)			
Planas	Ivykdyta	Būklė	Publikacijos tipas
Publikacija mokslo leidinyje, turinčiame cituojamumo rodiklį „Clarivate Analytics Web of Science“ duomenų bazėje. Planuojama publikacijos tema: mašininio mokymosi metodų, naudojamų kenkėjiškų programų generavimui ir obfuskavimui, tyrimas ir gautų rezultatų pateikimas.	Dautartas J., Kurasova O., Čypas J. R., Medvedev V. <i>Learning to Look Benign: Targeted Evasion of Malware Detectors via API Import Injection</i> . Rankraštis įteiktas žurnalui „IEEE Transactions on Information Forensics and Security“ (SCIE, IF 8.7; Q1); išankstinė versija paskelbta archyve „arXiv“ (arXiv:2605.18624).	Įteikta (2026 m. gegužės 20 d.)	Su citavimo rodikliu (CA WoS), SCIE, IF 8.7; Q1
Publikacija periodiniame recenzuojamame mokslo žurnale arba konferencijų medžiagoje.	Dautartas J., Čypas J. R., Kurasova O., Medvedev V. <i>Evasion of Malware Classifiers by Injecting Category-Specific Benign Features</i> . 2026 International Conference on Advances in Artificial Intelligence and Machine Learning (AAIML 2026). IEEE, 2026, p. 102–109. DOI: 10.1109/AAIML67890.2026.11498090.	Išspausdinta	Be citavimo rodiklio (recenzuojama konferencijos medžiaga)

Tyrimų objektas

Mašininio mokymosi algoritmai C2 sistemų agentams (kenkėjiškoms programoms) generuoti ir obfuskuoti.

Tikslas

Pasiūlyti ir ištirti naują kenkėjiškos programinės įrangos generavimo metodą, skirtą veiksmingai išnaudoti mašininiu mokymusi pagrįstų kenkėjiškų programų aptikimo sistemų pažeidžiamumą, siekiant padidinti kibernetinių grėsmių atsparumą ir pagerinti bendrą jų aptikimo tikslumą ir patikimumą.

Uždaviniai

- Atlikti išsamią analitinės literatūros apžvalgą, siekiant nustatyti esamus mašininio mokymosi metodus ir būdus, skirtus kenkėjiškoms programoms aptikti, klasifikuoti ir generuoti kenkėjiškas programas.
- Analizuoti antivirusinių ir galinių įrenginių aptikimo ir reagavimo (angl. Endpoint Detection and Response, EDR) sistemų kenkėjiškos programinės įrangos aptikimo mechanizmus, siekiant suprasti kenkėjiškos programinės įrangos klasifikavimo ypatybes.
- Įvertinti ir išskirti pagrindinius kenkėjiškos programinės įrangos požymius, turinčius įtakos jų aptikimui, siekiant pagerinti kenksmingų programų klasifikavimo tikslumą.
- Sukurti naują arba modifikuoti esamą kenkėjiškų programų (angl. Command and Control framework agents) generavimo metodą, kuriuo siekiama sukurti klaidinančius ir sunkiai aptinkamus C2 sistemų agentų pavyzdžius, galinčius klaidinti mašininio mokymosi pagrįstus kenkėjiškų programų aptikimo modelius.
- Įvertinti sugeneruotus kenkėjiškų programų pavyzdžius, naudojant įvairius kenkėjiškų programų aptikimo sprendimus ir mašininio mokymosi grindžiamas kibernetinio saugumo sistemas.
- Iširti esamas ir pasiūlyti naują strategiją, kaip padidinti mašininio mokymosi modelių atsparumą priešiškos atakoms.



Per pusmetį gauti rezultatai

- Atlikta vieno mėnesio mokslinė stažuotė Tenesio universitete Noksvilyje (JAV) (Liepos 2 – 30 d.), numatyta doktorantūros studijų plane.
- Straipsnis „Evasion of Malware Classifiers by Injecting Category-Specific Benign Features“ (J. Dautartas, J. R. Čypas, O. Kurasova, V. Medvedev) išspausdintas tarptautinės konferencijos AAIML 2026 straipsnių rinkinyje (IEEE, p. 102–109, DOI: 10.1109/AAIML67890.2026.11498090).
- Užbaigtas ir žurnalui „IEEE Transactions on Information Forensics and Security“ įteiktas straipsnis „Learning to Look Benign: Targeted Evasion of Malware Detectors via API Import Injection“ (J. Dautartas, O. Kurasova, J. R. Čypas, V. Medvedev), kuriame pristatomas sąlyginio variaciniu autoenkoderiu (CVAE) grįstas tikslinio priešiškojo išvengimo metodas ir jo eksperimentinis įvertinimas. Išankstinė straipsnio versija paskelbta atvirosios prieigos archyve „arXiv“ (arXiv:2605.18624).
- Tyrimuose naudojama šešių klasių duomenų aibė „WinAPI-AdvMal: A Six-Class Windows API Import Dataset for Adversarial Malware“ (3 799 Windows PE formato failai, aprašyti 2 713 dvejetainių Win32 API importo požymių paskelbta atvirosios prieigos saugykloje „Zenodo“ (DOI: 10.5281/zenodo.20208958).
- Pradėtas gynybinės krypties tyrimas, skirtas mašininio mokymosi modelių atsparumo priešiškomis atakoms didinimui.
- Kartu su bendraautoriais atliktas statinių kenkėjiškų programų klasifikatorių vertinimas pagal gerybinės programinės įrangos funkcines kategorijas. Rezultatai priimti skelbti konferencijos (CORE A* level) „ACM SIGSAC Conference on Computer and Communications Security“ (CCS '26, 2026 m. lapkričio 15–19 d., Haga, Nyderlandai) medžiagoje kaip stendinis pranešimas „Beyond a Single Benign Class: What Functional Categories of Benign Software Reveal About Static Malware Detection“ (V. Medvedev, J. Dautartas, J. R. Čypas, O. Kurasova) DOI: 10.1145/3830454.3846425.



From attack to defence

Adversarial robustness of WinAPI imports based malware classification

PART I — ATTACK

Develop adversarial attacks that evade malware detection

Add API imports to mislead the classifier. Use a CVAE and ensemble-guided target selection to generate evasive API-import modifications

„IEEE Transactions on Information Forensics and Security“ įteiktas straipsnis „Learning to Look Benign: Targeted Evasion of Malware Detectors via API Import Injection“ (J. Dautartas, O. Kurasova, J. R. Čypas, V. Medvedev)

Attacks guide
defence design

PART II — DEFENCE

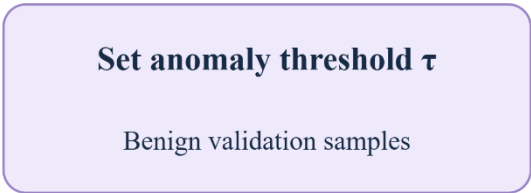
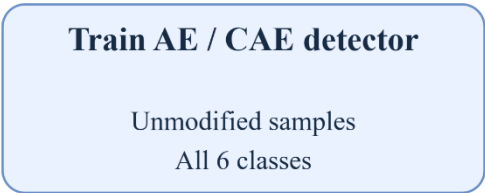
Develop and evaluate defences against adversarial malware evasion

Combine anomaly detection, feature reconstruction, and ensemble classification for adversarial defence

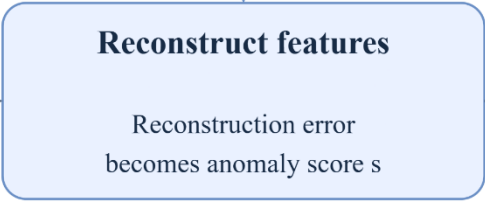
Parengti ir įteikti straipsnį apie priešiškų pavyzdžių aptikimą mokslo leidiniui, turinčiam cituojamumo rodiklį „Clarivate Web of Science“ duomenų bazėje. Planuojama teikti žurnalo „Engineering Applications of Artificial Intelligence“

Primary idea of adversarial malware detection

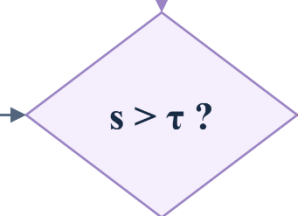
PREPARATION



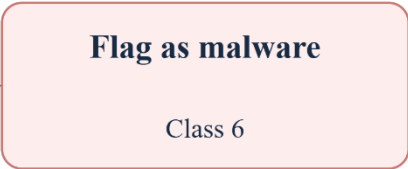
DETECTION



Score s

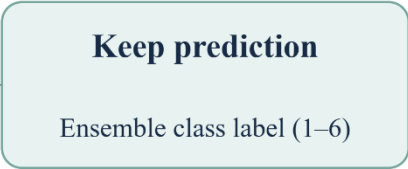


Threshold τ



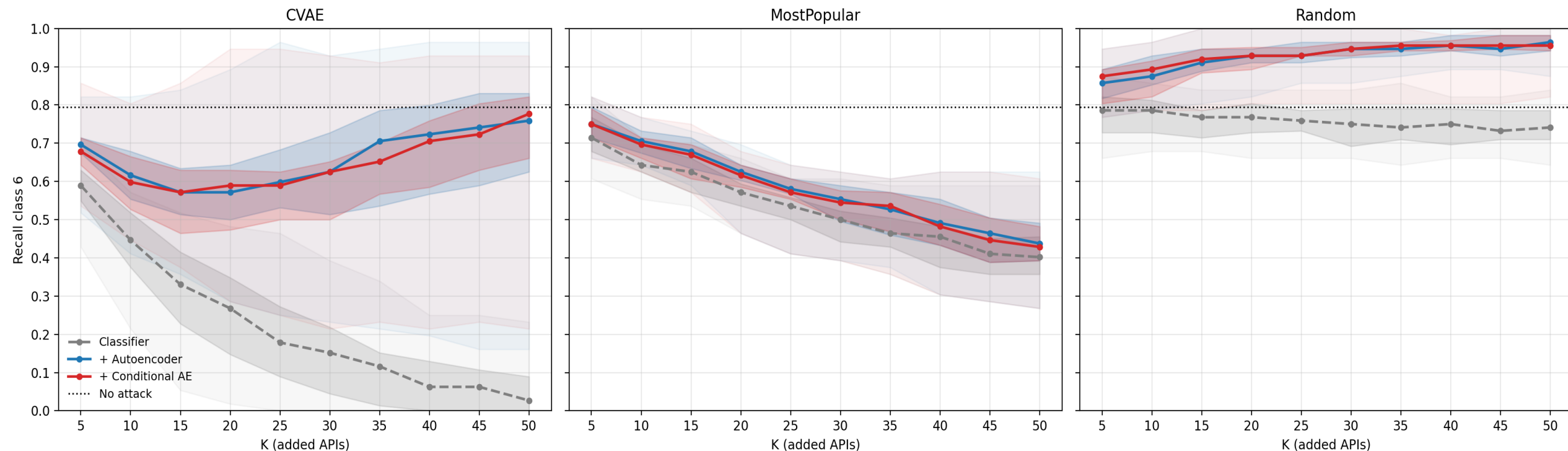
Anomaly

No anomaly



AE and CAE are alternative detectors. CAE keeps the lowest reconstruction error across the 6 class conditions.

Pirminiai eksperimentai siekiant padidinti atsparumą priešiškos atakoms naudojant klasikinį ir sąlyginį autoenkoderius. Lyginame skirtingų atakų poveikius klasifikatorių 6 klasės (malware) recall reikšmei. Pilka linija rodo kokia būtų recall reikšmė jeigu netaikytume autoenkoderių kaip papildomų apsaugos priemonių prieš pirminį klasifikatorių.



Malware klasės recall reikšmės prie skirtingų atakų

Attack	System	Recall (mean $\pm$ std)
None	Classifier	0.786 $\pm$ 0.057
	+ AE	0.806 $\pm$ 0.055
	+ CAE	0.804 $\pm$ 0.056
CVAE	Classifier	0.055 $\pm$ 0.075
	+ AE	0.729 $\pm$ 0.184
	+ CAE	0.692 $\pm$ 0.201
MostPopular	Classifier	0.412 $\pm$ 0.095
	+ AE	0.441 $\pm$ 0.101
	+ CAE	0.433 $\pm$ 0.099
Random	Classifier	0.742 $\pm$ 0.059
	+ AE	0.958 $\pm$ 0.031
	+ CAE	0.954 $\pm$ 0.041

Kito pusmečio darbo planas

- Užbaigti priešiškų pavyzdžių aptikimo eksperimentus. Palyginti autoenkoderiu ir sąlyginiu autoenkoderiu grįstus detektorius, įvertinti prisitaikančio atakuotojo (angl. adaptive attacker) scenarijų ir metodo ribotumus.
- Parengti ir įteikti straipsnį apie priešiškų pavyzdžių aptikimą mokslo leidiniui, turinčiam cituojamumo rodiklį „Clarivate Web of Science“ duomenų bazėje. Planuojama teikti žurnalo „Engineering Applications of Artificial Intelligence“, Elsevier, SCIE IF = 9, Q1 specialiajam numeriui „Human-Centered and Trustworthy AI for Cybersecurity in Cyber-Physical Systems“.
- Tęsti žurnalui „IEEE Transactions on Information Forensics and Security“ įteikto straipsnio publikavimo procesą. Atsižvelgti į recenzentų pastabas ir prireikus atlikti papildomus eksperimentus.
- Dalyvauti tarptautinėje mokslinėje konferencijoje ir pristatyti naujausius tyrimo rezultatus (pagal studijų planą – 2027 m. I ketvirtį). Konferencija: 2027 2nd International Conference on Advances in Artificial Intelligence and Machine Learning (AAIML 2027), kovo 29–31 d., 2027 m. (Japonija).
- Patikslinti disertacijos tikslą, uždavinius, tyrimo metodiką ir ginamuosius teiginius.
- Papildyti ir atnaujinti analitinę disertacijos dalį ir tęsti teorinės bei eksperimentinės dalių rengimą, įtraukiant CVAE grįsto metodo ir priešiškų pavyzdžių aptikimo tyrimų rezultatus.
- Plėtoti stažuotės metu užmegztus mokslinius ryšius su Tenesio universiteto ir Oak Ridžo nacionalinės laboratorijos tyrėjų grupėmis.



KLAUSIMAI

Juozas Dautartas
juozas.dautartas@mif.stud.vu.lt