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 pirmas pusmetis

Doktorantas: Juozas Dautartas
Vadovas: Dr. Viktor Medvedev

Studijų planas ir jo vykdymo suvestinė

Vilniaus universitetas

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**		**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 (I pusmetis)		
Planas	**Įvykdyta**	**Būklė**
-	-	-

Dalyvavimas konferencijose 2025/2026 (I pusmetis)		
Planas	**Įvykdyta**	**Konferencijos tipas**
Dalyvavimas tarptautinėje mokslinėje konferencijoje.	Pranešėjas: Juozas Dautartas Pranešimo pavadinimas: Evasion of Malware Classifiers by Injecting Category-Specific Benign Features Konferencijos pavadinimas: 2026 International Conference on Advances in Artificial Intelligence and Machine Learning (AAIML 2026) Data: 2026 m. kovo 20–22 d. Vieta: Tokijas, Japonija (Chuo University)	Tarptautinė
	Pranešėjas: Juozas Dautartas Pranešimo pavadinimas: Feature level deception or when malware wears a mask Konferencijos pavadinimas: 16th conference on data analysis methods for software systems (DAMSS) Data: 2025 m. lapkričio 27–29 d. Vieta: Druskininkai, Lietuva	Nacionalinė

Publikacijos 2025/2026 (I pusmetis)			
Planas	**Įvykdyta**	**Būklė**	**Publikacijos tipas**

Tyrimų objektas

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

Tikslas

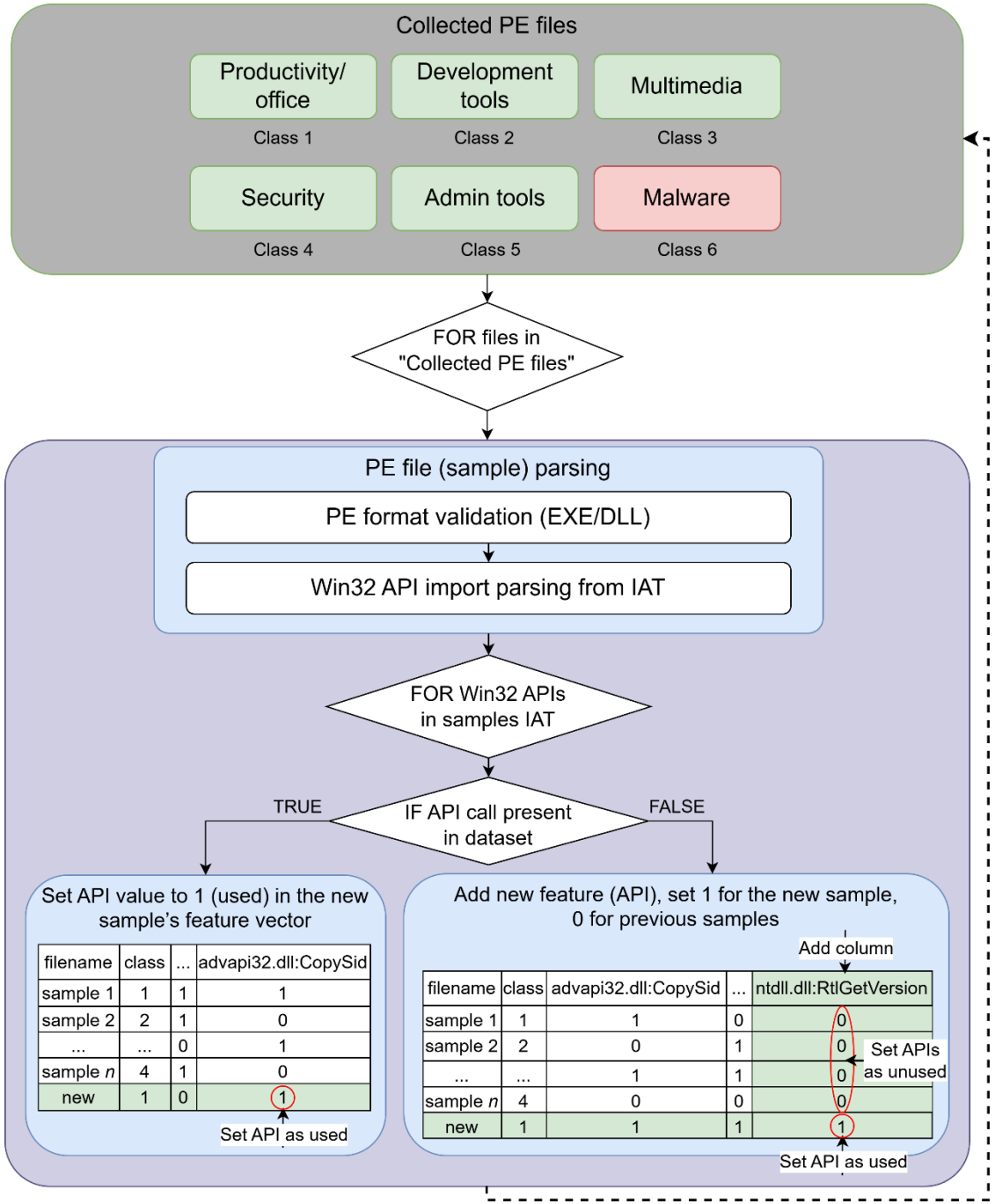
Pasiūlyti ir ištirti naują kenkėjiškos programinės įrangos generavimo metodą, skirtą veiksmingai išnaudoti mašininiu mokymusi pagrįstų kenkėjiškų programų aptikimo sistemų pažeidžiamumą, siekiant padidinti kibernetinių grėsmių atsparumą ir pagerinti bendrą jų aptikimo tikslumą ir patikimumą.

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 Responce, 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 mašininiu mokymusi pagrįstus kenkėjiškų programų aptikimo modelius.
- Įvertinti sugeneruotus kenkėjiškų programų pavyzdžius, naudojant įvairius kenkėjiškų programų aptikimo sprendimus ir mašininiu mokymusi grindžiamas kibernetinio saugumo sistemas.
- Ištirti esamas ir pasiūlyti naują strategiją, kaip padidinti mašininio mokymosi modelių atsparumą priešiškoms atakoms.

Per pusmetį gauti rezultatai

- Ankstesnė duomenų aibė (852 PE failai, 3 gerybinės klasės) išplėsta. Sudaryta kelių kategorijų Windows PE formato failų duomenų aibė, apimanti 6 klases (5 gerybinių programų funkcinės kategorijos ir viena kenkėjiškos programinės įrangos klasė; 3 799 pavyzdžiai, 5 713 dvejetainių požymių). Duomenų aibė paskelbta Nacionaliniame atviros prieigos mokslinių tyrimų duomenų archyve: Budžys A., Dautartas J., Medvedev V., Kurasova O., Čypas J. R., Jomantas H. (2025). WinAPI-4C-AdvMal: Windows API features for adversarial malware. MIDAS. DOI: 10.18279/MIDAS.265677.
- Remiantis ankstesnio etapo rezultatais ir išplėsta duomenų aibe, pasiūlytas ir sukurtas sąlyginio variacinio autoenkoderio(angl. CVAE) pagrįstas tikslinio priešiškojo išvengimo metodas su apribojimu, leidžiantis įterpti tik naujus importus nepakeičiant originalios kenkėjiškos programos funkcionalumo.
- Buvo sukurtas ir įgyvendintas keturių klasifikatorių ansamblis, kuris sprendimus priima vertindamas tiek pirminius požymius (angl. raw features), tiek gautus įterpinius (angl. embeddings). Šiems įterpiniams generuoti atskirai pasiūlytas neuroninis tinklas, naudojant lanko ilgio (angl. ArcFace) ir prižiūrimo kontrastinio mokymosi (angl. SupCon) nuostolių funkcijas, kurios leidžia efektyviau sugrupuoti panašius pavyzdžius ir pagerinti skirtingų klasių atskyrimą.
- Parengta publikacija mokslo leidiniui, turinčiam cituojamumo rodiklį Clarivate Web of Science duomenų bazėje. Planuojama įteikti žurnalui „IEEE Transactions on Information Forensics and Security“ (Q1 (computer science, theory & methods in SCIE), 8 IF). Planuojama publikacijos tema: Mašininio mokymosi metodų, naudojamų kenkėjiškų programų generavimui ir obfuskavimui tyrimas ir gautų rezultatų pateikimas.

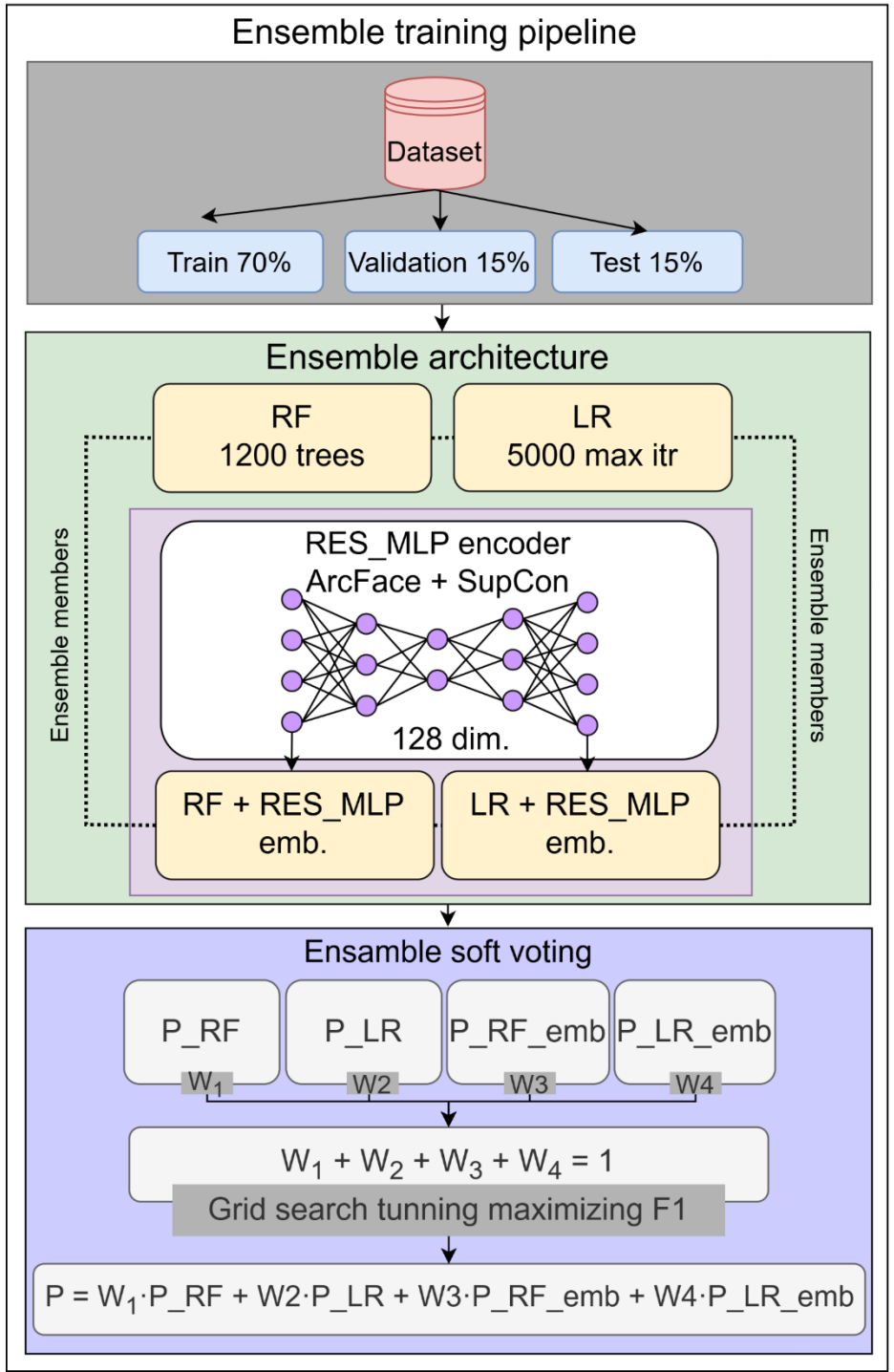


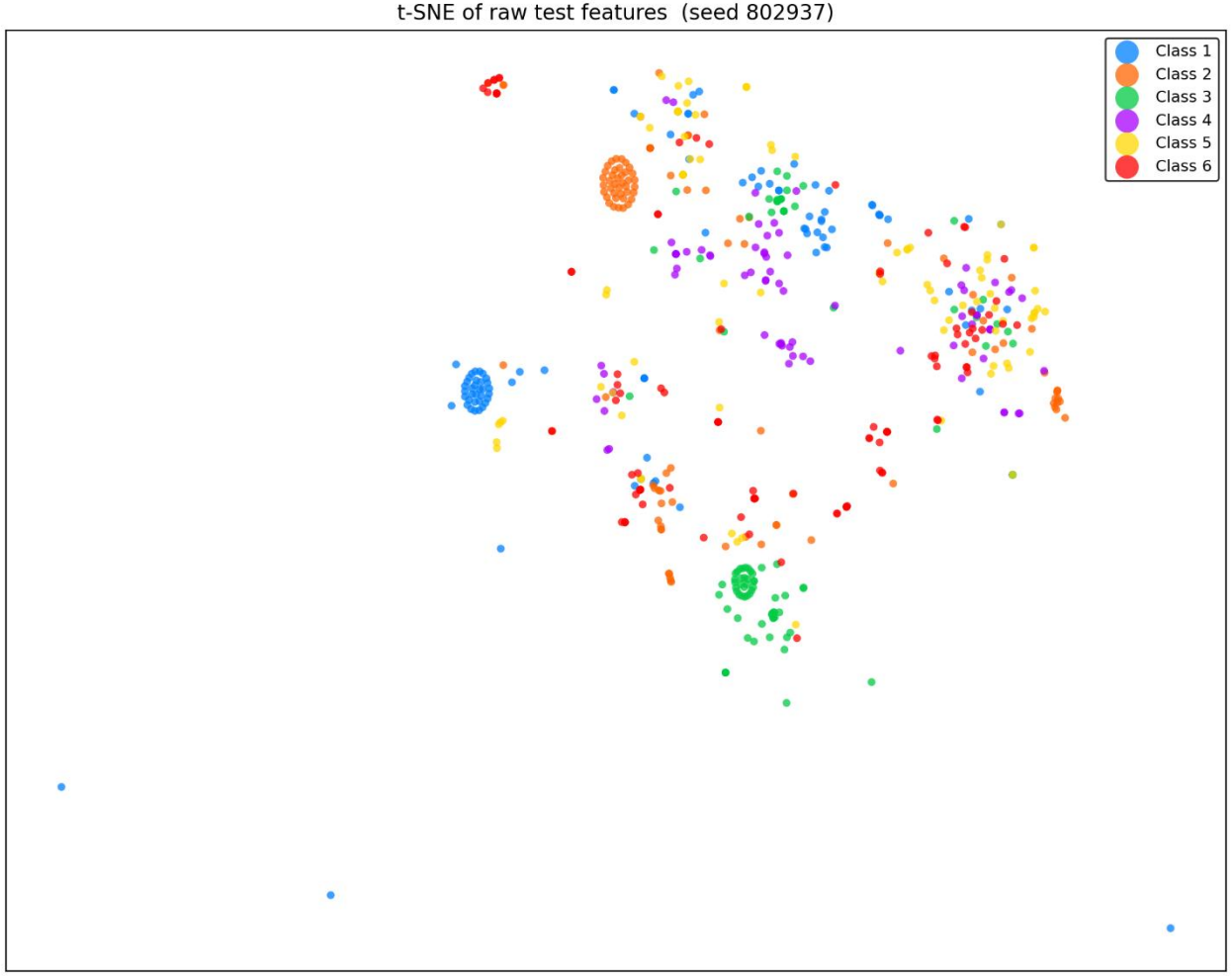
Vilniaus universitetas

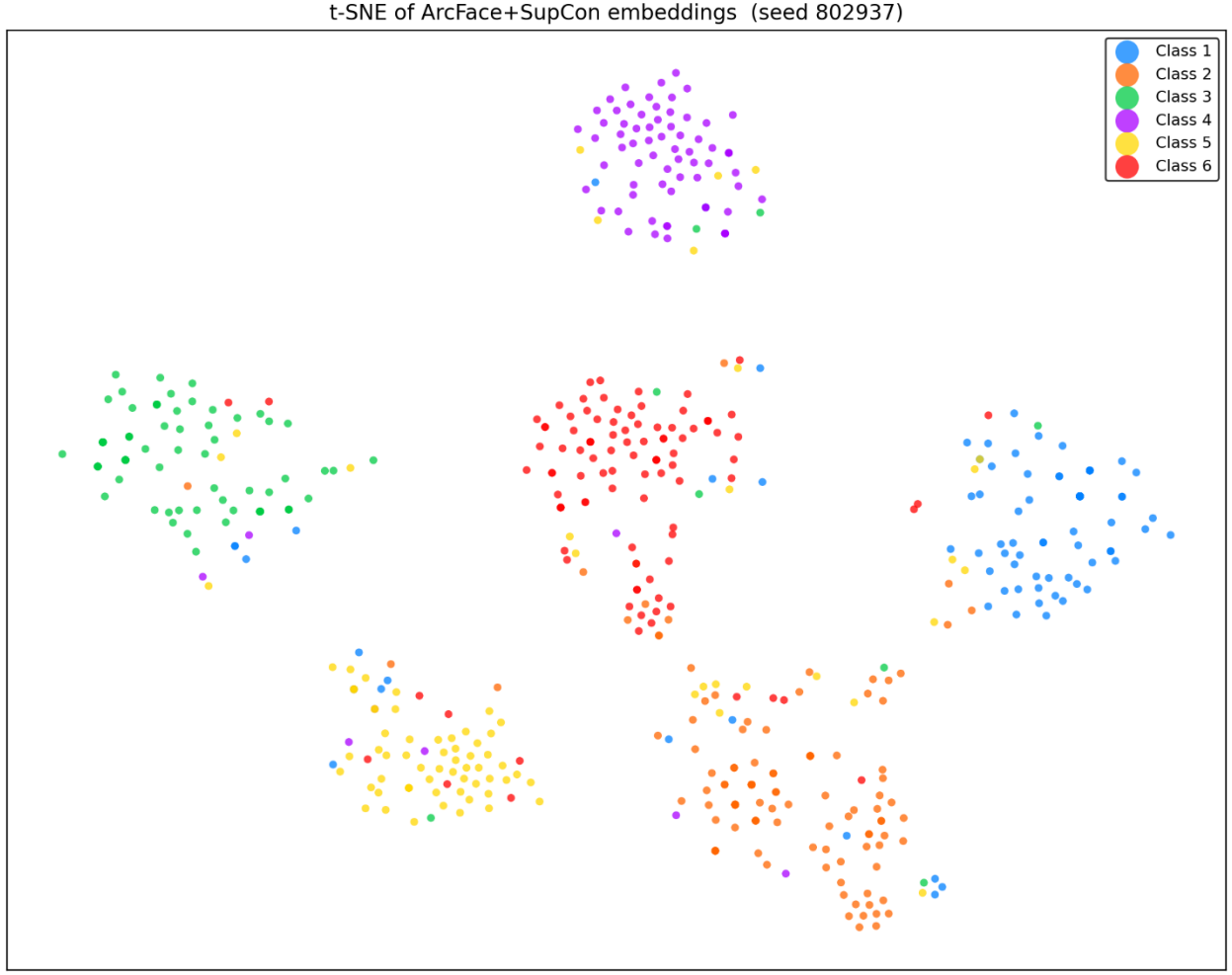
Duomenų rinkimo eiga

Ansambliai klasifikatoriai

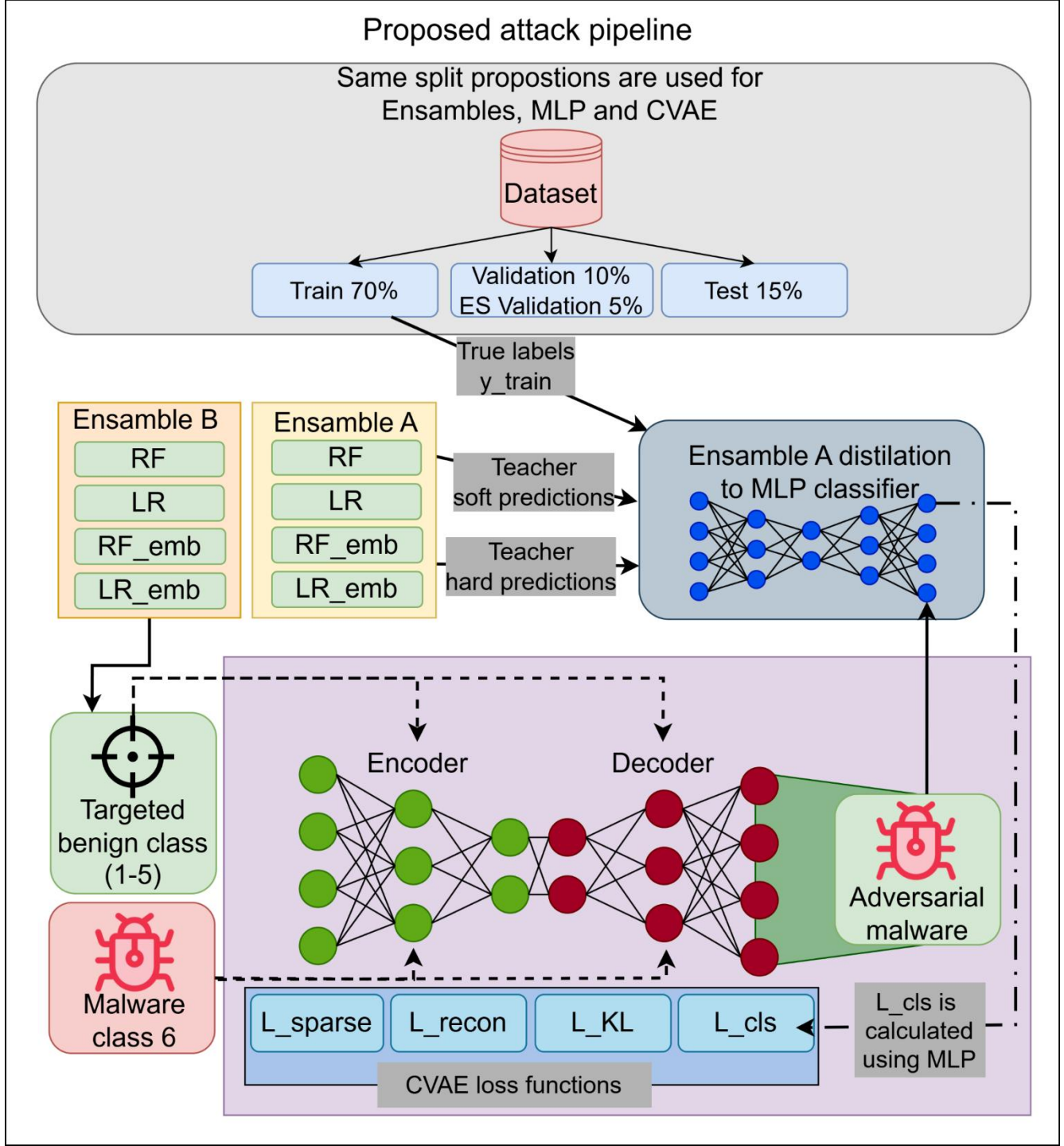
- Apmokome du ansamblius:
  - **Ansamblis A** – apmokomas naudojant visas klases (įskaitant ir virusų klasę).
  - **Ansamblis B** – apmokomas be virusų klasės.

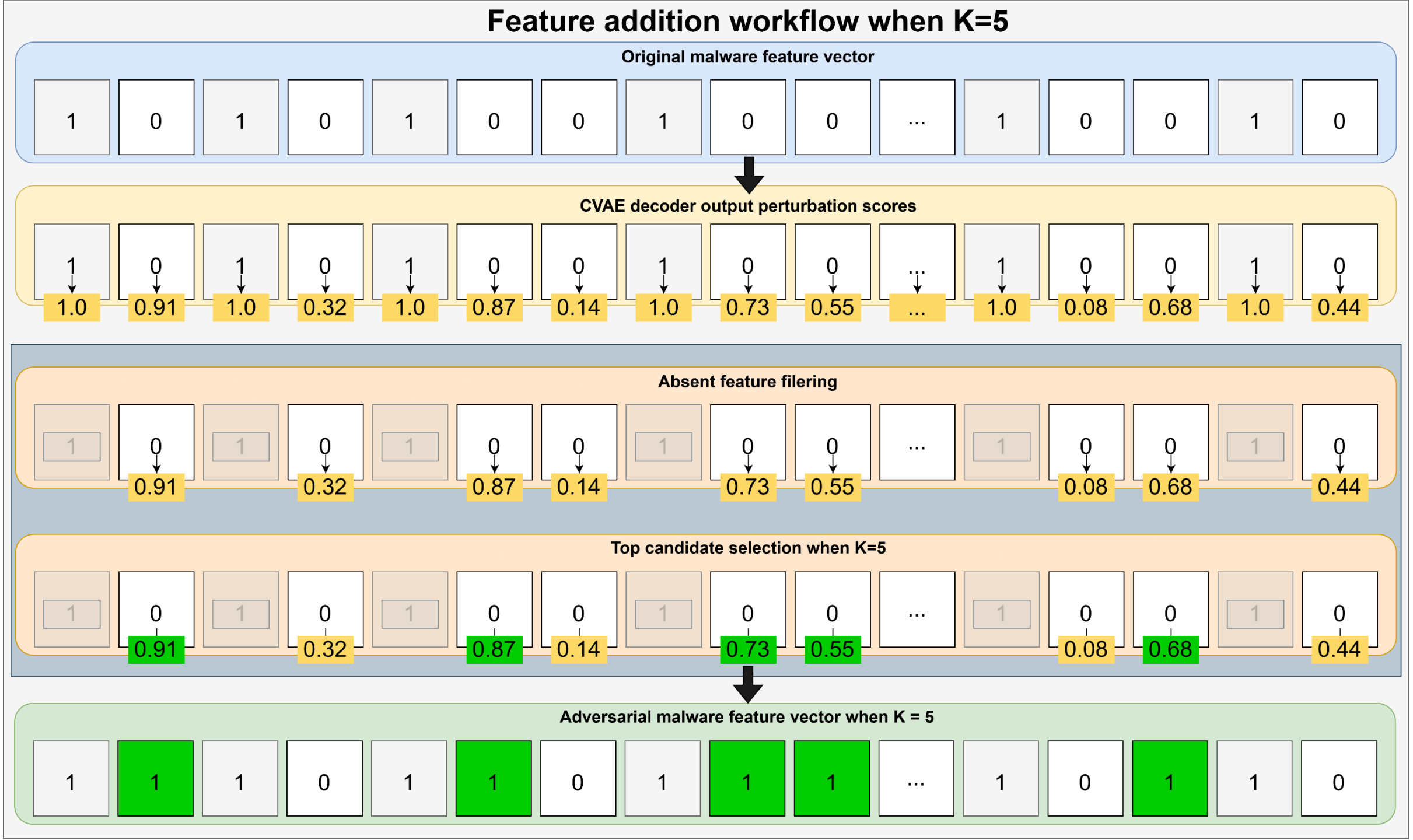



t-SNE of raw test features (seed 802937)
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6


t-SNE of ArcFace+SupCon embeddings (seed 802937)
Class 1
Class 2
Class 3
Class 4
Class 5
Class 6

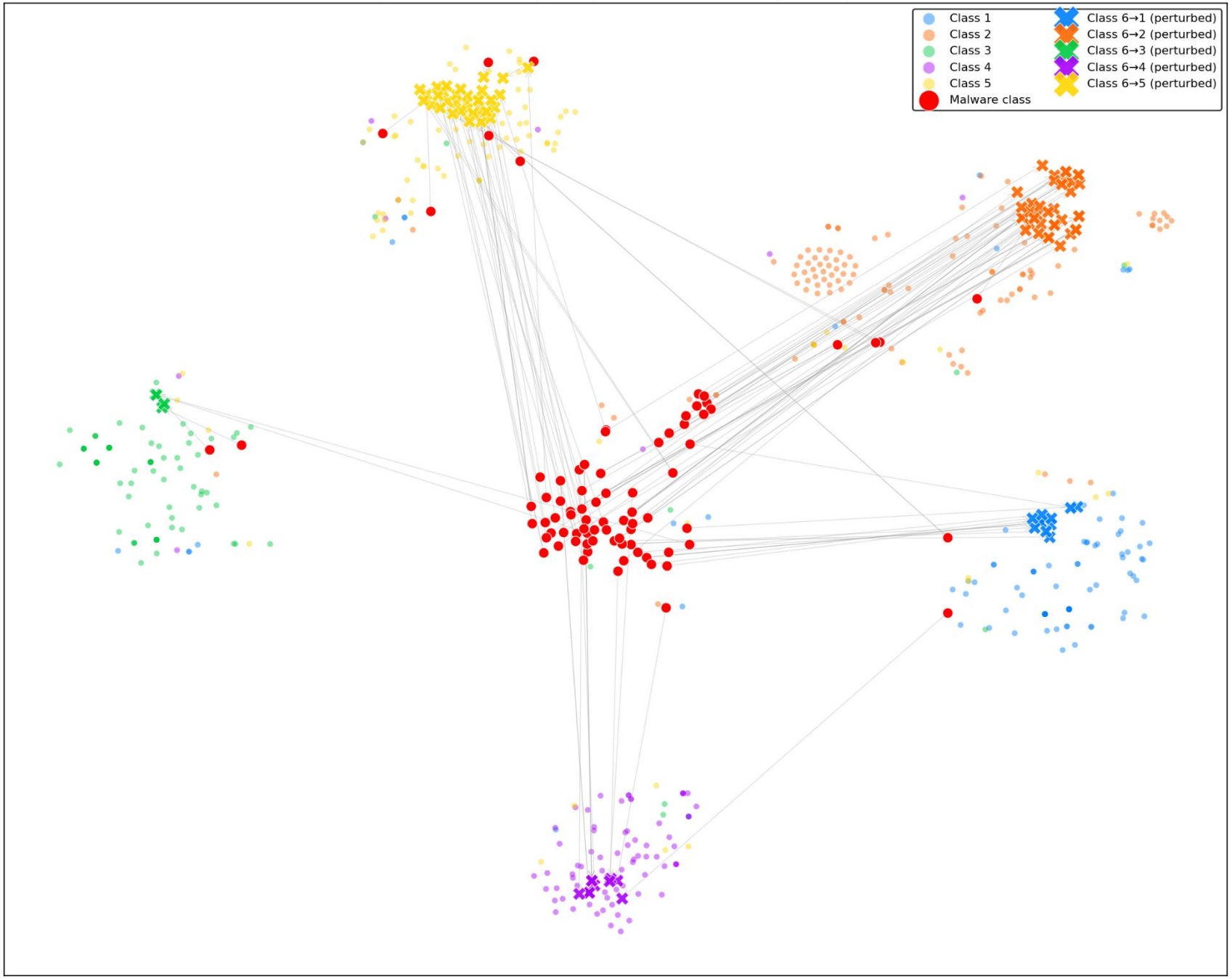
CVAE apmokymas ir ataka

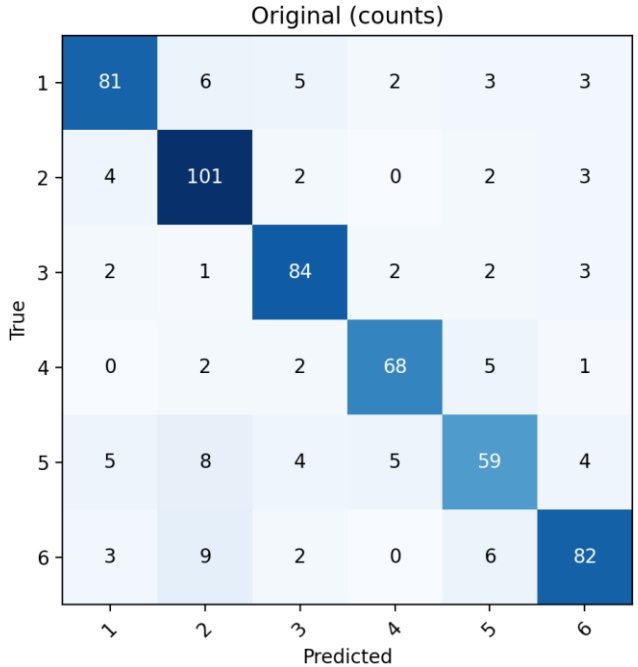


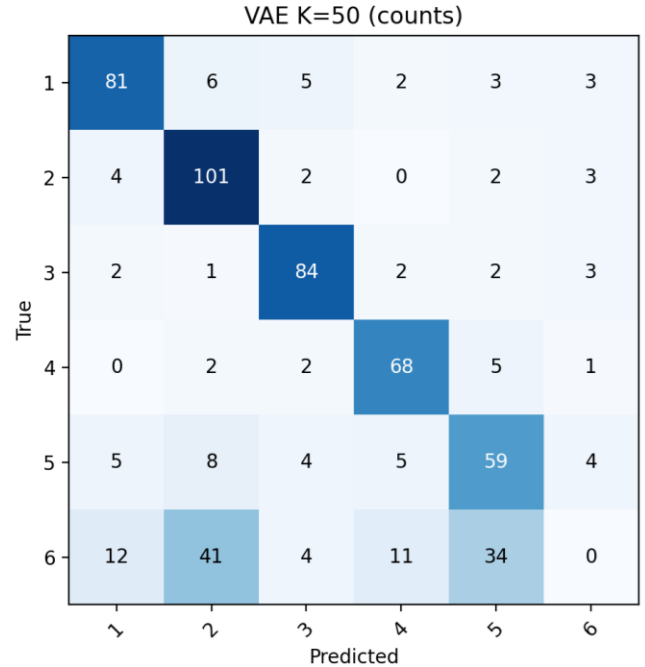

Feature addition workflow when K=5
Original malware feature vector
1 0 1 0 1 0 0 1 0 0 ... 1 0 0 1 0
CVAE decoder output perturbation scores
1 0 1 0 1 0 0 1 0 0 ... 1 0 0 1 0
1.0 0.91 1.0 0.32 1.0 0.87 0.14 1.0 0.73 0.55 ... 1.0 0.08 0.68 1.0 0.44
Absent feature filering
1 0 1 0 1 0 0 1 0 0 ... 1 0 0 1 0
0.91 0.32 0.87 0.14 0.73 0.55 0.08 0.68 0.44
Top candidate selection when K=5
1 0 1 0 1 0 0 1 0 0 ... 1 0 0 1 0
0.91 0.32 0.87 0.14 0.73 0.55 0.08 0.68 0.44
Adversarial malware feature vector when K = 5
1 1 1 0 1 1 0 1 1 1 ... 1 0 1 1 0

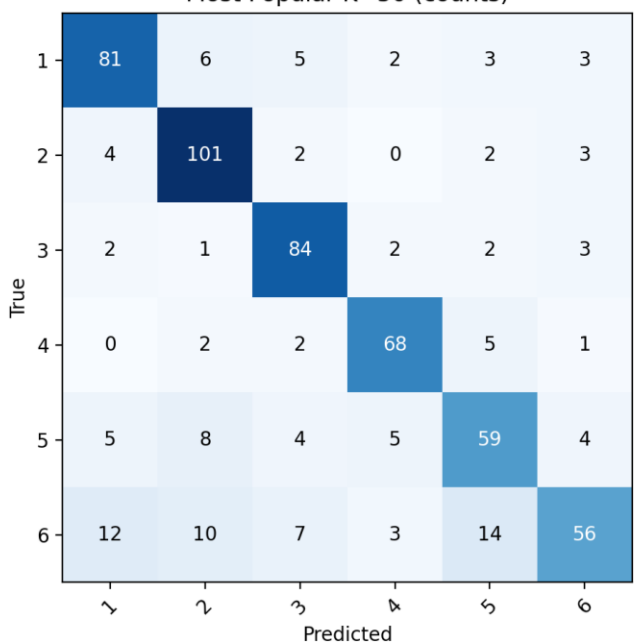
t-SNE: class-6 movement after CVAE perturbation (K=50, seed 802937)

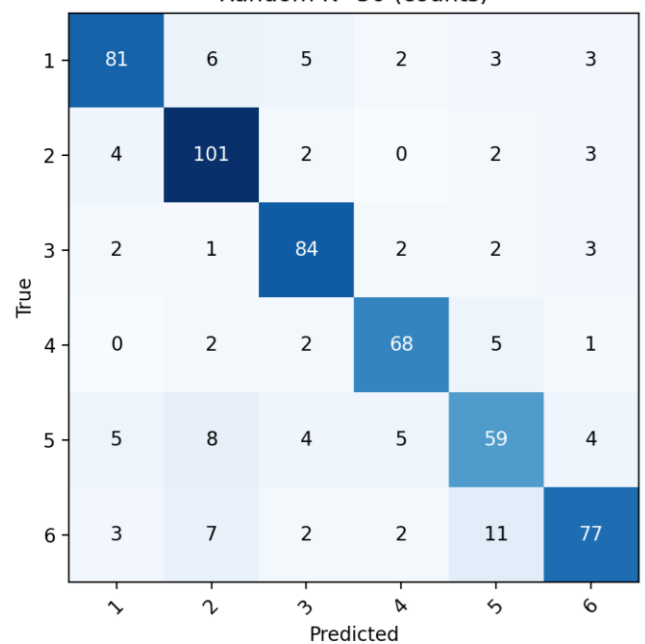
● = Malware class | X = perturbed (colour = target class)




Original (counts)
True
Predicted
1 2 3 4 5 6
81 6 5 2 3 3
4 101 2 0 2 3
2 1 84 2 2 3
0 2 2 68 5 1
5 8 4 5 59 4
3 9 2 0 6 82


VAE K=50 (counts)
True
Predicted
1 2 3 4 5 6
81 6 5 2 3 3
4 101 2 0 2 3
2 1 84 2 2 3
0 2 2 68 5 1
5 8 4 5 59 4
12 41 4 11 34 0


Most Popular K=50 (counts)
True
Predicted
1 2 3 4 5 6
81 6 5 2 3 3
4 101 2 0 2 3
2 1 84 2 2 3
0 2 2 68 5 1
5 8 4 5 59 4
12 10 7 3 14 56


Random K=50 (counts)
True
Predicted
1 2 3 4 5 6
81 6 5 2 3 3
4 101 2 0 2 3
2 1 84 2 2 3
0 2 2 68 5 1
5 8 4 5 59 4
3 7 2 2 11 77

	CVAE				**MostPopular**				**Random**			
K	UER	TSR	CTS	Rec._6	UER	TSR	CTS	Rec._6	UER	TSR	CTS	Rec._6
0	–	–	–	$0.87_{\pm 05}$	–	–	–	$0.87_{\pm 05}$	–	–	–	$0.87_{\pm 05}$
5	$\mathbf{0.25}_{\pm 08}$	$\mathbf{0.24}_{\pm 09}$	$\mathbf{0.96}_{\pm 06}$	$\mathbf{0.75}_{\pm 08}$	$0.16_{\pm 05}$	$0.14_{\pm 05}$	$0.85_{\pm 12}$	$0.84_{\pm 05}$	$0.13_{\pm 04}$	$0.11_{\pm 04}$	$0.87_{\pm 10}$	$0.87_{\pm 04}$
10	$\mathbf{0.41}_{\pm 18}$	$\mathbf{0.40}_{\pm 18}$	$\mathbf{0.96}_{\pm 05}$	$\mathbf{0.59}_{\pm 18}$	$0.23_{\pm 06}$	$0.20_{\pm 07}$	$0.87_{\pm 09}$	$0.77_{\pm 06}$	$0.13_{\pm 04}$	$0.11_{\pm 04}$	$0.86_{\pm 11}$	$0.87_{\pm 04}$
15	$\mathbf{0.56}_{\pm 20}$	$\mathbf{0.56}_{\pm 20}$	$\mathbf{0.98}_{\pm 03}$	$\mathbf{0.44}_{\pm 20}$	$0.26_{\pm 05}$	$0.23_{\pm 05}$	$0.87_{\pm 09}$	$0.74_{\pm 05}$	$0.12_{\pm 04}$	$0.10_{\pm 04}$	$0.83_{\pm 12}$	$0.88_{\pm 04}$
20	$\mathbf{0.70}_{\pm 18}$	$\mathbf{0.70}_{\pm 18}$	$\mathbf{0.99}_{\pm 01}$	$\mathbf{0.30}_{\pm 18}$	$0.31_{\pm 05}$	$0.28_{\pm 06}$	$0.91_{\pm 06}$	$0.69_{\pm 05}$	$0.13_{\pm 05}$	$0.11_{\pm 04}$	$0.84_{\pm 12}$	$0.87_{\pm 05}$
25	$\mathbf{0.79}_{\pm 17}$	$\mathbf{0.78}_{\pm 17}$	$\mathbf{0.99}_{\pm 01}$	$\mathbf{0.21}_{\pm 17}$	$0.37_{\pm 06}$	$0.34_{\pm 07}$	$0.93_{\pm 06}$	$0.63_{\pm 06}$	$0.13_{\pm 05}$	$0.11_{\pm 03}$	$0.82_{\pm 14}$	$0.87_{\pm 05}$
30	$\mathbf{0.84}_{\pm 17}$	$\mathbf{0.83}_{\pm 16}$	$\mathbf{0.99}_{\pm 01}$	$\mathbf{0.16}_{\pm 17}$	$0.39_{\pm 07}$	$0.36_{\pm 07}$	$0.94_{\pm 06}$	$0.61_{\pm 07}$	$0.13_{\pm 05}$	$0.10_{\pm 03}$	$0.80_{\pm 14}$	$0.87_{\pm 05}$
50	$\mathbf{0.95}_{\pm 12}$	$\mathbf{0.94}_{\pm 12}$	$\mathbf{0.99}_{\pm 01}$	$\mathbf{0.05}_{\pm 12}$	$0.50_{\pm 10}$	$0.50_{\pm 10}$	$0.98_{\pm 03}$	$0.50_{\pm 10}$	$0.15_{\pm 05}$	$0.11_{\pm 03}$	$0.78_{\pm 15}$	$0.85_{\pm 05}$

Kito pusmečio darbo planas

- Pabaigti ir pateikti parengtą publikaciją mokslo leidiniui, turinčiam cituojamumo rodiklį Clarivate Web of Science duomenų bazėje (planuojama įteikti „IEEE Transactions on Information Forensics and Security“).
- Tęsti empirinį tyrimą ir tobulinti pasiūlytą CVAE grįstą metodą, tirti jo apibendrinamumą.
- Toliau tikslinti analitinę literatūros apžvalgą ir pradėti nuosekliai rengti daktaro disertacijos teorinę bei eksperimentinę dalis, remiantis atliktais CVAE grįsto metodo ir klasifikatorių ansamblių tyrimais.
- Planuojama išvykti į stažuotę užsienio mokslo ir studijų institucijoje (planuojamas stažuotės vieta: Tenesio universitetas, JAV).
- Pradėti rengti antrą publikaciją mokslo leidiniui, turinčiam cituojamumo rodiklį Clarivate Web of Science duomenų bazėje. Publikacija bus skirta pasiūlyto metodo tobulinimui ir mašininiu mokymusi grindžiamų kibernetinio saugumo sistemų atsparumo priešiškoms atakoms vertinimui.
- Atlikti papildomus eksperimentus, apibendrinti pasiektus rezultatus ir suformuluoti tarpines išvadas, kurios bus integruojamos į rengiamą disertaciją bei publikacijas.

KLAUSIMAI

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