Vilniaus universitetas
Duomenų mokslo ir skaitmeninių technologijų institutas

INFORMATIKOS KRYPTIES DOKTORANTŲ ATESTACINĖ KONFERENCIJA
VEIKLOS ATASKAITA UŽ 2022 M. SPALIO 1 D. – 2023 M. KOVO 21 D.

ANOMALINIŲ ĮVYKIŲ IDENTIFIKAVIMAS IR JŲ UŽKARDYMAS KOMPIUTERIŲ TINKLUOSE TAIKANT MAŠININIO MOKYMOSI METODUS

DOKT. ARNOLDAS BUDŽYS – INFORMATIKA N 009
STUDIJŲ METAI: III
DARBO VADOVAS: DR. VIKTOR MEDVEDEV
DOKTORANTŪROS PRADŽIOS IR PABAIGOS METAI: 2020–2024

2022–2023 m.

STUDIJŲ PLANAS IR JO VYKDYMO SUVESTINĖ

Studijų metai	Egzaminai [1]	
	Planas	Įvykdyta
I (2020/2021)	1	1
II (2021/2022)	3	3
III (2022/2023)		
IV (2023/2024)		
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 (2020/2021)										
II (2021/2022)	1	1	1	1				1	0	
III (2022/2023)	**1**	**1+1***	**0**	**1**	**1**			**1** (skola iš II metų)	**1+1***	Priimta +Įteikta
IV (2023/2024)					1					
Iš viso:	2	2	1	2	2			1	1	

*HCI2023 – priimta (CA WoS, Springer), CISTI2023 – įteikta (CA WoS, IEEE)

ATASKAITINIS STUDIJŲ PUSMETIS (III: 2022/2023 – I pusmetis)

Dalyvavimas konferencijose 2022/2023 (I pusmetis)		
Planas	Įvykdyta	Konferencijos tipas
Data Analysis Methods for Software Systems 2022 m. gruodžio 1–3 d., Druskininkai	Intrusion detection based on keystroke biometrics and siamese neural networks Data Analysis Methods for Software Systems 2022 m. gruodžio 1-3 d., Druskininkai Autoriai: Arnoldas Budžys, Olga Kurasova, Viktor Medvedev	Nacionalinė

Publikacijos 2022/2023 (I pusmetis)			
Planas	Įvykdyta	Būklė	Publikacijos tipas
Data Analysis Methods for Software Systems 2022 m. gruodžio 1–3 d., Druskininkai	Budžys, A.; Kurasova, O.; Medvedev, V. Intrusion detection based on keystroke biometrics and siamese neural networks // DAMSS: 13th conference on data analysis methods for software systems, Druskininkai, Lithuania, December 1–3, 202. Vilnius : Vilnius University Press, 2022. ISBN 9786090707944. eISBN 9786090707951. p. 13. (Vilnius University Proceedings, eISSN 2669-0233 ; vol. 31). DOI: 10.15388/DAMSS.13.2022.	Publikuotas	Konferencijos pranešimo tezės

ATASKAITINIS STUDIJŲ PUSMETIS (III: 2022/2023 – I pusmetis)

Publikacijos 2022/2023 (I pusmetis)			
Planas	Įvykdyta	Būklė	Publikacijos tipas
25th International Conference On Human-Computer Interaction 2023 m. liepos 23–28 d. Kopenhaga, Danijos Karalystė	Budžys, A., Kurasova, O., and Medvedev, V., "Behavioral Biometrics Authentication Using Siamese Neural Networks", in HCI for Cybersecurity, Privacy and Trust 5th International Conference, HCI-CPT 2023, Held as Part of the 25th HCI International Conference, HCI2023, 2023, pp. 1–14 (in press).	Priimta: 2023.02.10	CA WoS duomenų bazėje be cituojamumo rodiklio
18th Iberian Conference on Information Systems and Technologies 2023 m. birželio 20–23 d. Aveiro, Portugalija	Medvedev, V., Budžys, A., & Kurasova, O. Enhancing Keystroke Biometric Authentication Using Deep Learning Techniques. In 2023 18th Iberian Conference on Information Systems and Technologies (CISTI), 2023, pp. 1-6. IEEE. (Submitted, under review).	Įteikta: 2023.02.24	CA WoS duomenų bazėje be cituojamumo rodiklio

DOKTORANTŪROS STUDIJŲ PASIEKIMAI

Dalyvavimas tarptautinėse konferencijose	
1.	Budžys, A., Kurasova, O., and Medvedev, V., „Deep learning-based prevention of insider threats using user behavioral keystroke biometrics“, 32nd European Conference on Operational Research (EURO XXXII)], Espoo, Finland, July 3-6, 2022.
2.	Budžys, A., Kurasova, O., and Medvedev, V., “Behavioral Biometrics Authentication Using Siamese Neural Networks”, in HCI for Cybersecurity, Privacy and Trust 5th International Conference, HCI-CPT 2023, Held as Part of the 25th HCI International Conference, HCI2023, 2023, pp. 1–14 (in press).

Doktorantūros mokslinių tyrimų ir disertacijos rengimo etapai

Darbo pavadinimas	Atlikimo terminai	Pastabos
Mokslinių tyrimų disertacijos tema apžvalga ir analizė (Lietuvoje ir užsienyje):	2020 m. spalio mėn. – 2021 m. rugsėjo mėn.	Atlikus literatūros analizę pavyko identifikuoti problemos sprendimo būdus panaudojant dirbtinius neuroninius tinklus.
Mokslinio tyrimo vykdymas: 2.1. Tyrimo metodikos sudarymas: 2.1.1. Tyrimo metodikos iškeltiems uždaviniams spręsti parinkimas; 2.1.2. Teorinio ir empirinio tyrimų suplanavimas pagal pasirinktą metodiką. 2.2. **Teorinis tyrimas:** 2.2.1. Mašininio mokymosi metodų, naudojamų kompiuterių tinkluose įsilaužimų prevencijai, tyrimas. 2.2.2. 2.3. **Empirinis tyrimas:** **2.3.1. Sudarytų metodų pritaikymas praktinių uždavinių sprendimui.** **2.3.2. Gautų duomenų analizė, rezultatų apibendrinimas, išvadų parengimas.**	2021 m. spalio mėn. – 2022 m. sausio mėn. 2022 m. vasario mėn. – 2022 m. rugsėjo mėn. **2022 m. spalio mėn. – 2023 m. rugsėjo mėn.**	Pasiūlytas naujas metodas, kuriame duomenys transformuojami į vaizdinį pavidalą. Naujo metodo palyginimas su kitais literatūroje naudojamais metodais naudojant Siamo neuroninius tinklus (angl. Siamese Neural Network, SNN). Gauti eksperimentinio tyrimo rezultatai lyginami tarpusavyje. Tyrimai rodo, jog konvertavus skaitines reikšmes į vaizdus galima pagerinti vartotojų klasifikavimo rezultatą, lyginant su mašininio mokymosi algoritmais, bei klasikiniais dirbtiniais neuroniniais tinklais kuomet klasifikavimui naudojami skaitiniai duomenys.

Doktorantūros mokslinių tyrimų ir disertacijos rengimo etapai

Atskirų daktaro disertacijos dalių (tyrimo metodikos, rezultatų, ginamų teiginių, išvadų, ir kt.) parengimas: 3.1. Tikslų, uždavinių, tyrimo metodikos, ginamųjų teiginių patikslinimas; 3.2. Analitinės disertacijos dalies parengimas; 3.3. Teorinės disertacijos dalies parengimas; 3.4. Eksperimentinės disertacijos dalies parengimas; 3.5. Bendrųjų išvadų formulavimas.	2023 m. spalio mėn. – 2024 m. gegužės mėn.	
Daktaro disertacijos parengimas ir svarstymas padalinyje	2024 m. birželio mėn.	
Daktaro disertacijos gynimas	2024 m. rugsėjo mėn.	

Disertacijos tema, tyrimo objektai ir tikslas

Preliminari disertacijos tema:

- Anomalinių įvykių identifikavimas ir jų užkardymas kompiuterių tinkluose taikant mašininio mokymosi metodus.

Tyrimo objektai:

- vartotojo sugeneruoti klaviatūros, pelės biometriniai duomenys, bei mašininio mokymosi metodų taikymas anomalinių įvykių identifikavimui ir neteisėtų veiksmų užkardymui.

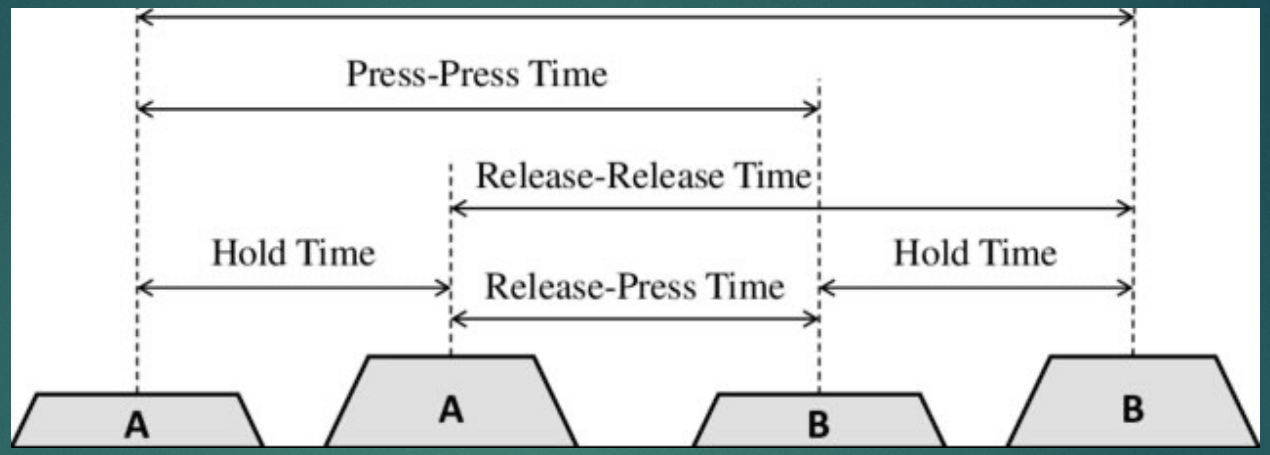
Tikslas:

- pasiūlyti metodiką sistemos vartotojui autentifikuoti pagal jo biometrinius elgsenos duomenis siekiant užkardyti insaiderio veiklą bei apsaugoti sistemą nuo jo neteisėtų veiksmų.

Tyrimo uždaviniai

- Atlikti išsamią literatūros analitinę apžvalgą, siekiant identifikuoti tinkamus metodus anomalinių įvykių identifikavimui ir insaiderio užkardymui kompiuterių tinkluose;
- Atlikti skirtingų mašininio mokymosi metodų, skirtų anomalinių įvykių identifikavimui ir insaiderio užkardymui kompiuterių tinkluose, analizę ir tyrimą;
- Sukurti metodiką, apimančią mašininiu mokymusi grįstus algoritmus, sistemos vartotojui autentifikuoti pagal jo biometrinius elgsenos duomenis;
- Įvertinti sukurtos metodikos efektyvumą realaus laiko duomenims atliekant eksperimentinius tyrimus;
- Atlikti gautų rezultatų analizę: rezultatų apibendrinimas, išvadų parengimas.

Klaviatūros biometriniai duomenys



1 pav. Klaviatūros biometrinių duomenų surinkimas[1]

0.1491	0.3979	0.2488	0.1069	0.1674	0.0605	0.1169	0.2212	0.1043	0.1417	1.1885
1.0468	0.1146	1.6055	1.4909	0.1067	0.759	0.6523	0.1016	0.2136	0.112	0.1349
0.1484	0.0135	0.0932	0.3515	0.2583	0.1338	0.3509	0.2171	0.0742		

References: 1. Harilal, A., Toffalini, F., Castellanos, J., Guarnizo, J., Homoliak, I., & Ochoa, M. (2017, October). Twos: A dataset of malicious insider threat behavior based on a gamified competition. In *Proceedings of the 2017 International Workshop on Managing Insider Security Threats* (pp. 45-56).

Duomenų aibės

- Carnegie Mellon University (CMU) dataset[2]
- GREYC dataset[3]
- Bei Hang dataset[4]
- Aalto University dataset[5]

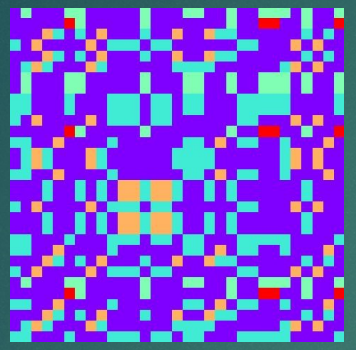
References: **2.** Killourhy, K. S., & Maxion, R. A. (2009, June). Comparing anomaly-detection algorithms for keystroke dynamics. In *2009 IEEE/IFIP International Conference on Dependable Systems & Networks* (pp. 125-134). IEEE. **3.** Giot, R., El-Abed, M., & Rosenberger, C. (2009, September). Greyc keystroke: a benchmark for keystroke dynamics biometric systems. In *2009 IEEE 3rd International Conference on Biometrics: Theory, Applications, and Systems* (pp. 1-6). IEEE**. 4.** Li, Y., Zhang, B., Cao, Y., Zhao, S., Gao, Y., & Liu, J. (2011, October). Study on the BeiHang keystroke dynamics database. In 2011 International Joint Conference on Biometrics (IJCB) (pp. 1-5). IEEE. **5.** *Dhakal, V., Feit, A. M., Kristensson, P. O., & Oulasvirta, A. (2018, April). Observations on typing from 136 million keystrokes. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-12).*

Siūloma metodika

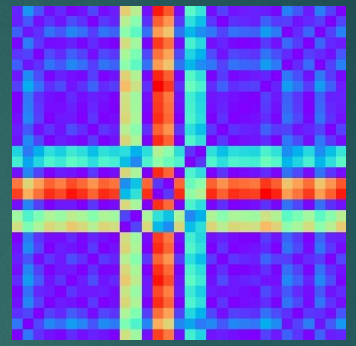
- Sistemos vartotojui autentifikuojantis sistemoje jo įvesto slaptažodžio laiko žymos iš skaitinio pavidalo konvertuojamos į vaizdinius, kurie yra lyginami su jau duomenų bazėje esančiu vartotojo šablonu.
- Prisijungus prie sistemos, vartotojo klaviatūros biometrika yra toliau stebima ir konvertuojama į vaizdinius ir kaskart tikrinama ar pirminiame etape autentifikavęsis vartotojas toliau naudojasi sesija.
- Esant abejonėms, dirbtinis neuroninis tinklas inicijuoja sistemos užrakinimą ir vartotojas turi iš naujo autentifikuotis pirminiame (statiniame) etape (žr. 3 paveikslą).

No Image to Image

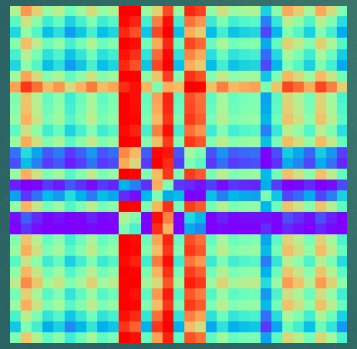
1. Markov Transition Field (MTF)[6]



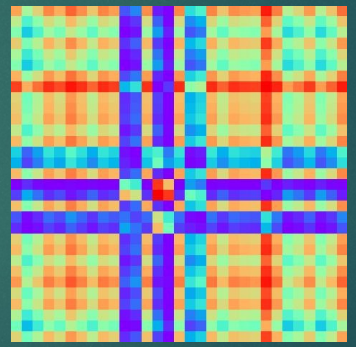
2. Reccurence Plots (RPs)[7]



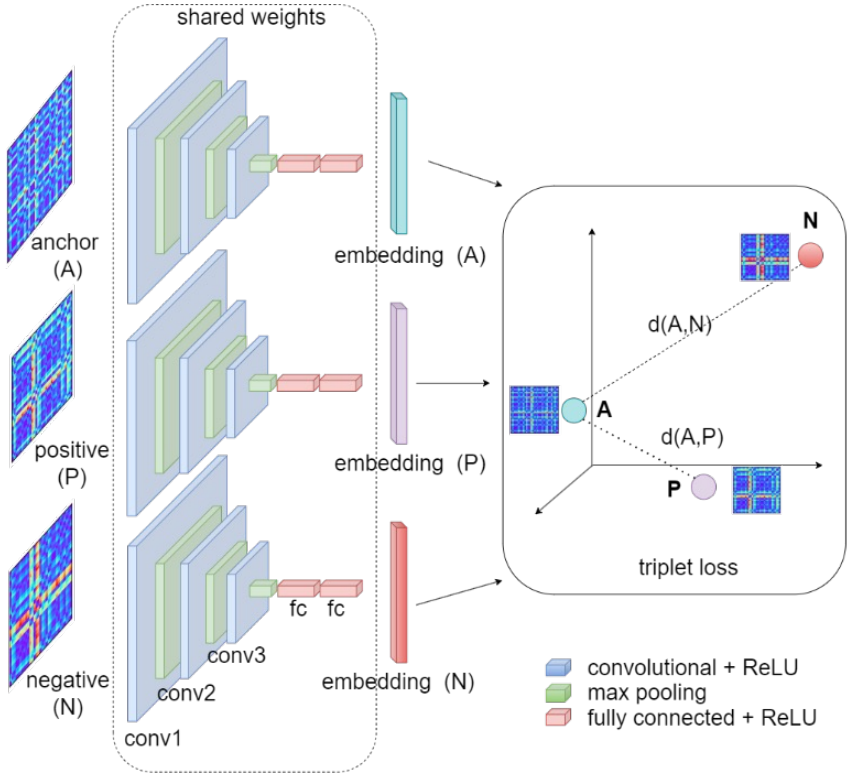
3. Gramian Angular Difference Field (GADF)[8]



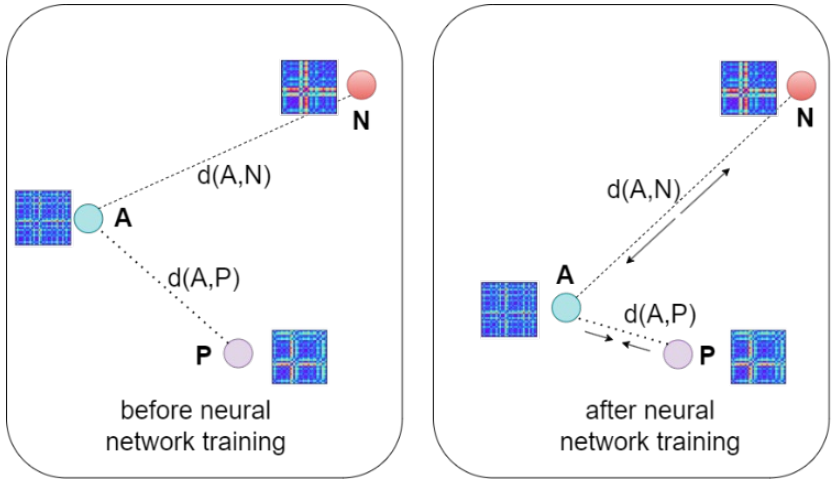
4. Gramian Angular Summation Field (GASF)[8]



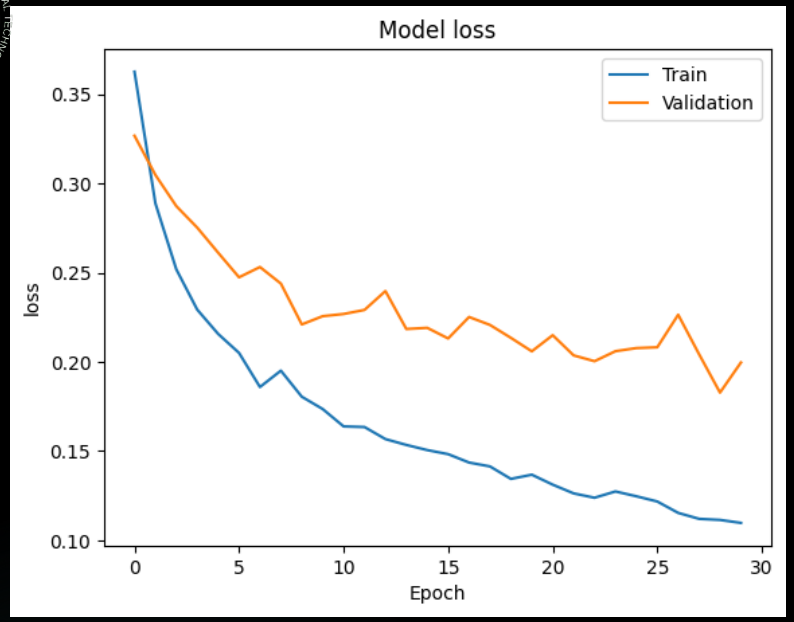
References: **6**. Rue, H., & Held, L. (2005). *Gaussian Markov random fields: theory and applications*. Chapman and Hall/CRC. **7.** Marwan, N., Romano, M. C., Thiel, M., & Kurths, J. (2007). Recurrence plots for the analysis of complex systems. *Physics reports*, *438*(5-6), 237-329. **8.** Wang, Z., & Oates, T. (2015, April). Encoding time series as images for visual inspection and classification using tiled convolutional neural networks. In Workshops at the twenty-ninth AAAI conference on artificial intelligence. **9.** Moustakidis, S., & Karlsson, P. (2020). A novel feature extraction methodology using Siamese convolutional neural networks for intrusion detection. *Cybersecurity*, *3*(1), 1-13.

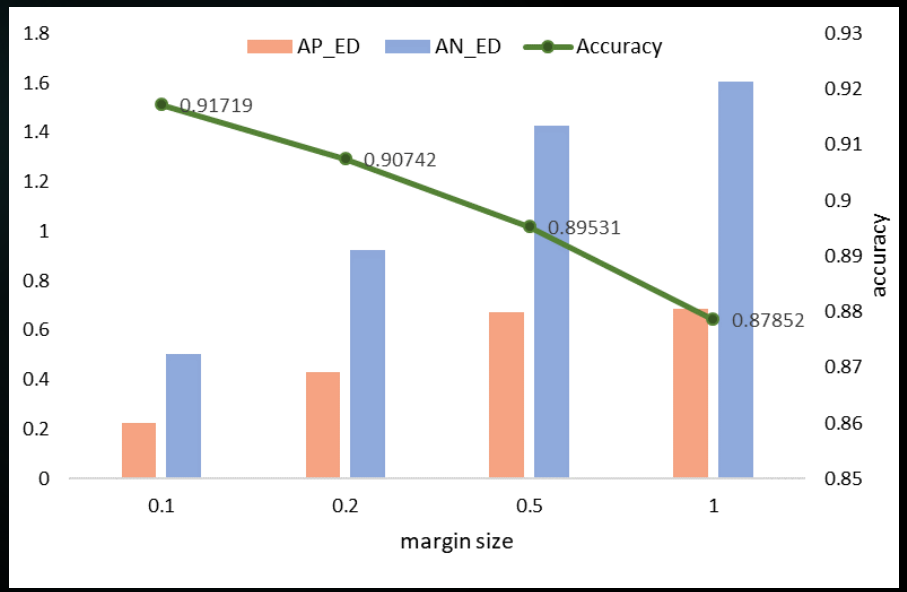


Example of Siamese neural network with a triplet loss function

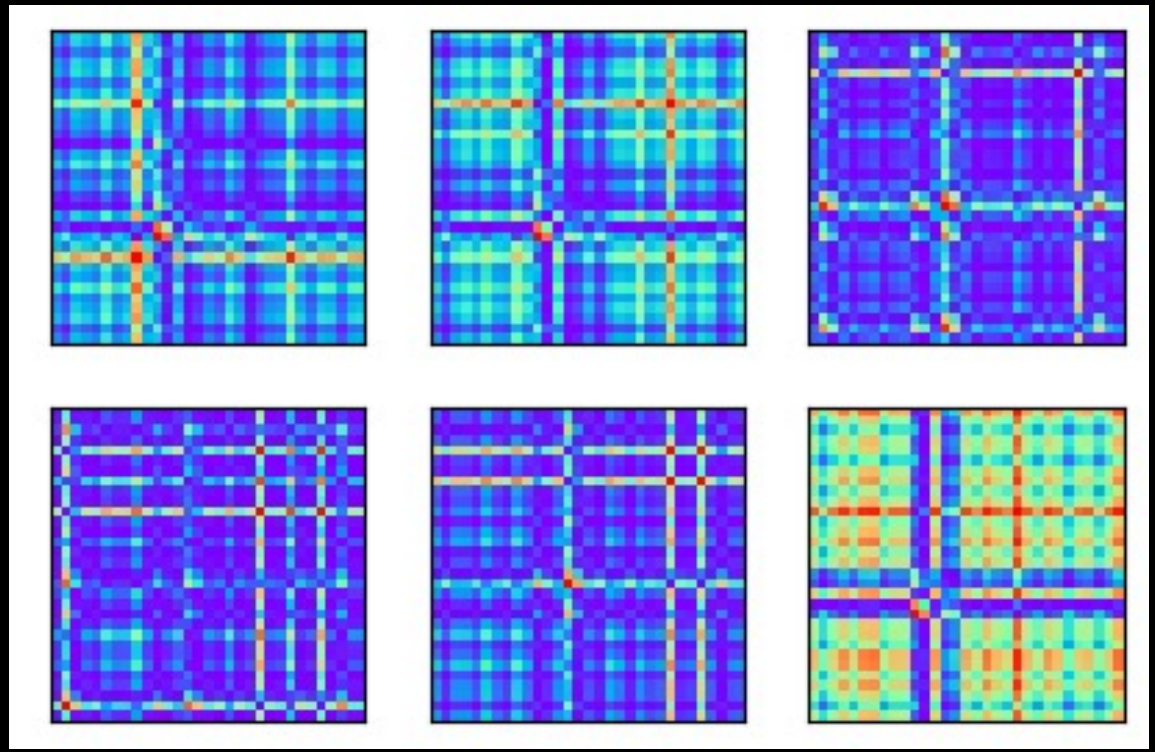


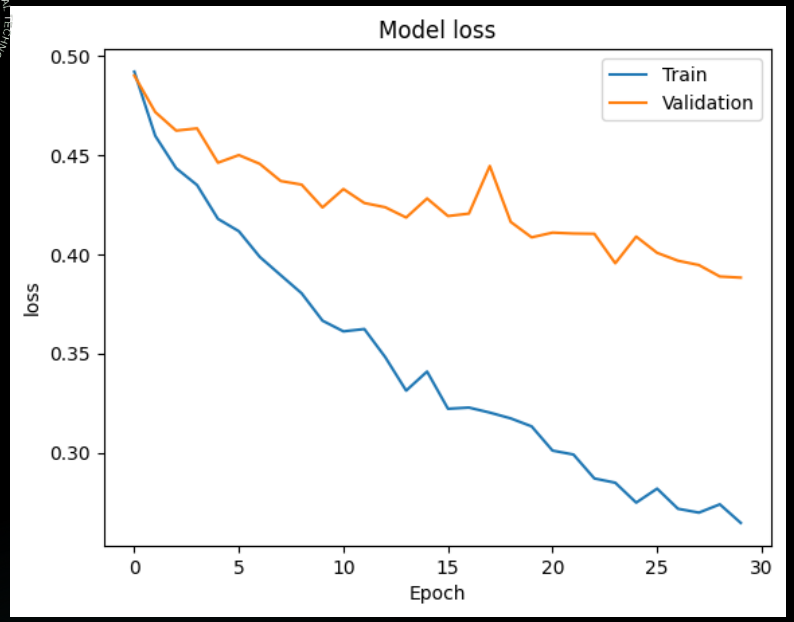
Example of a triplet before and after training a Siamese neural network.

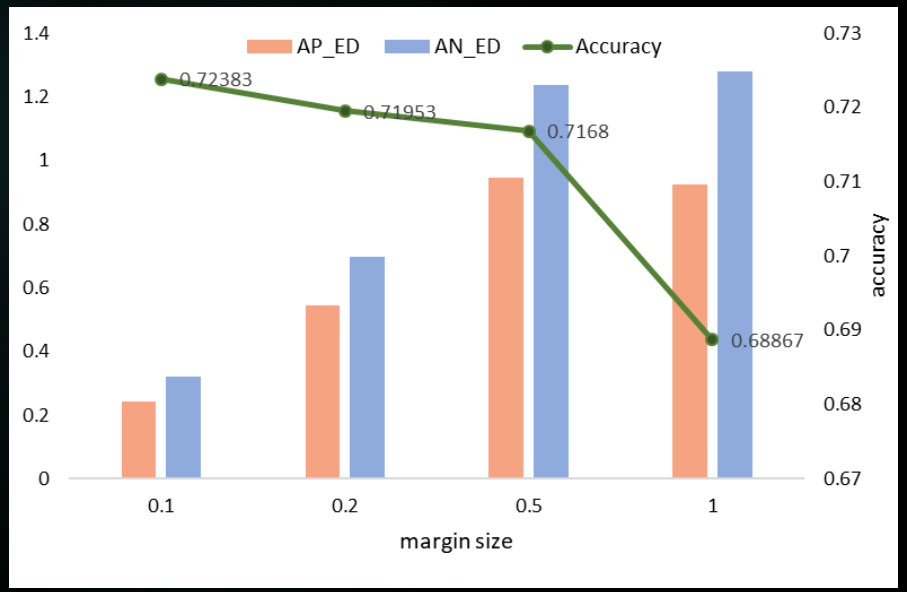




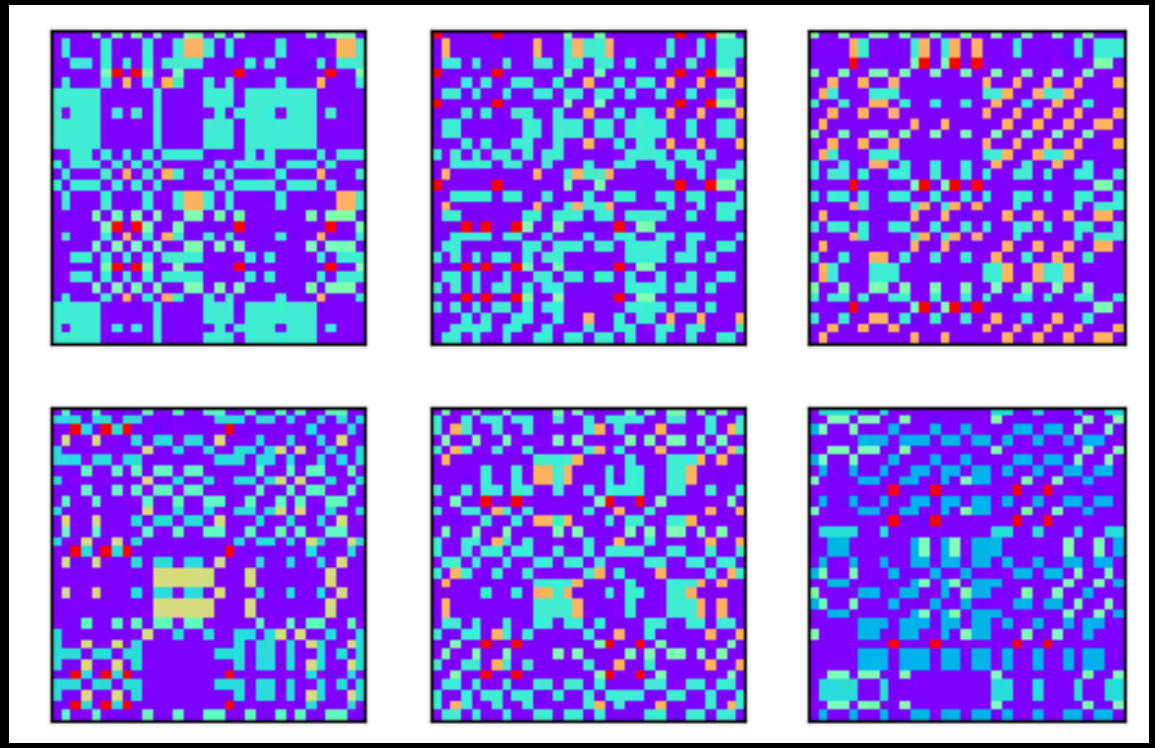
Gramian Angular Summation Field

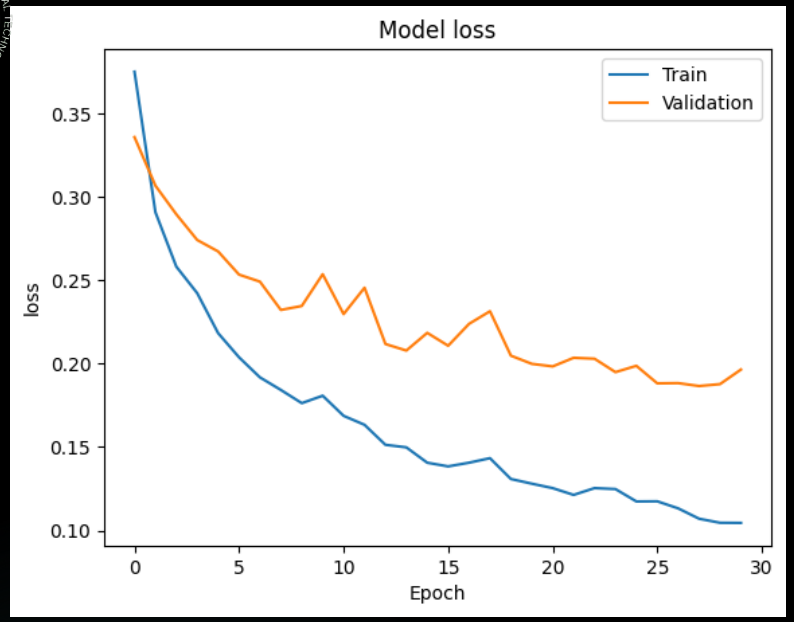


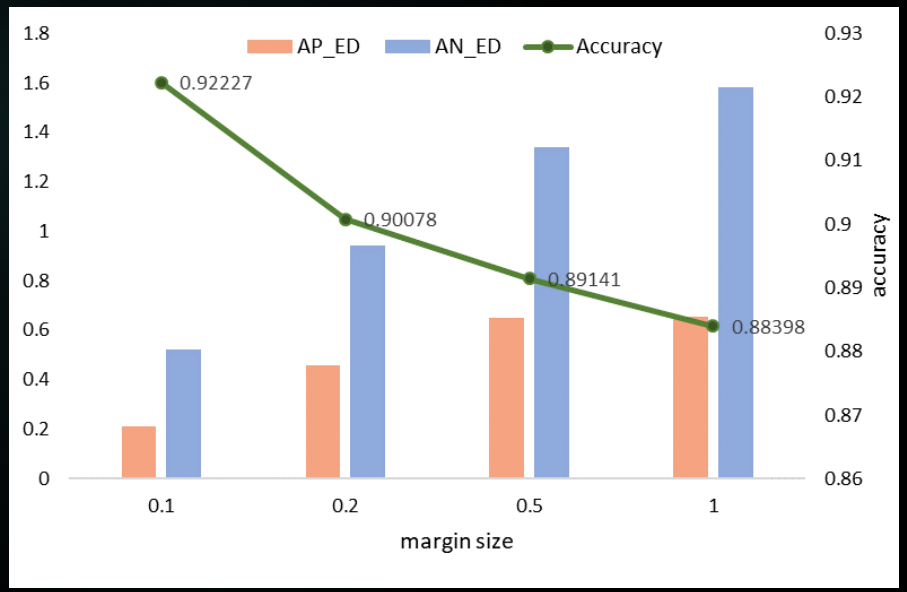




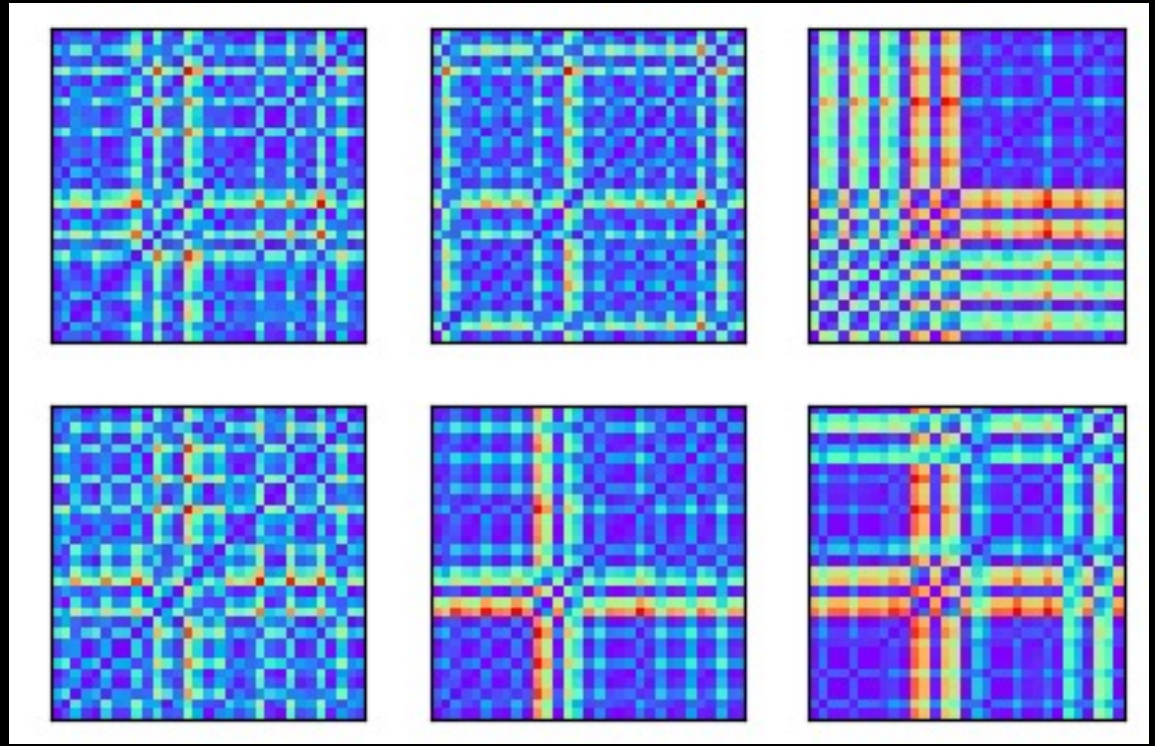
Markov Transition Field



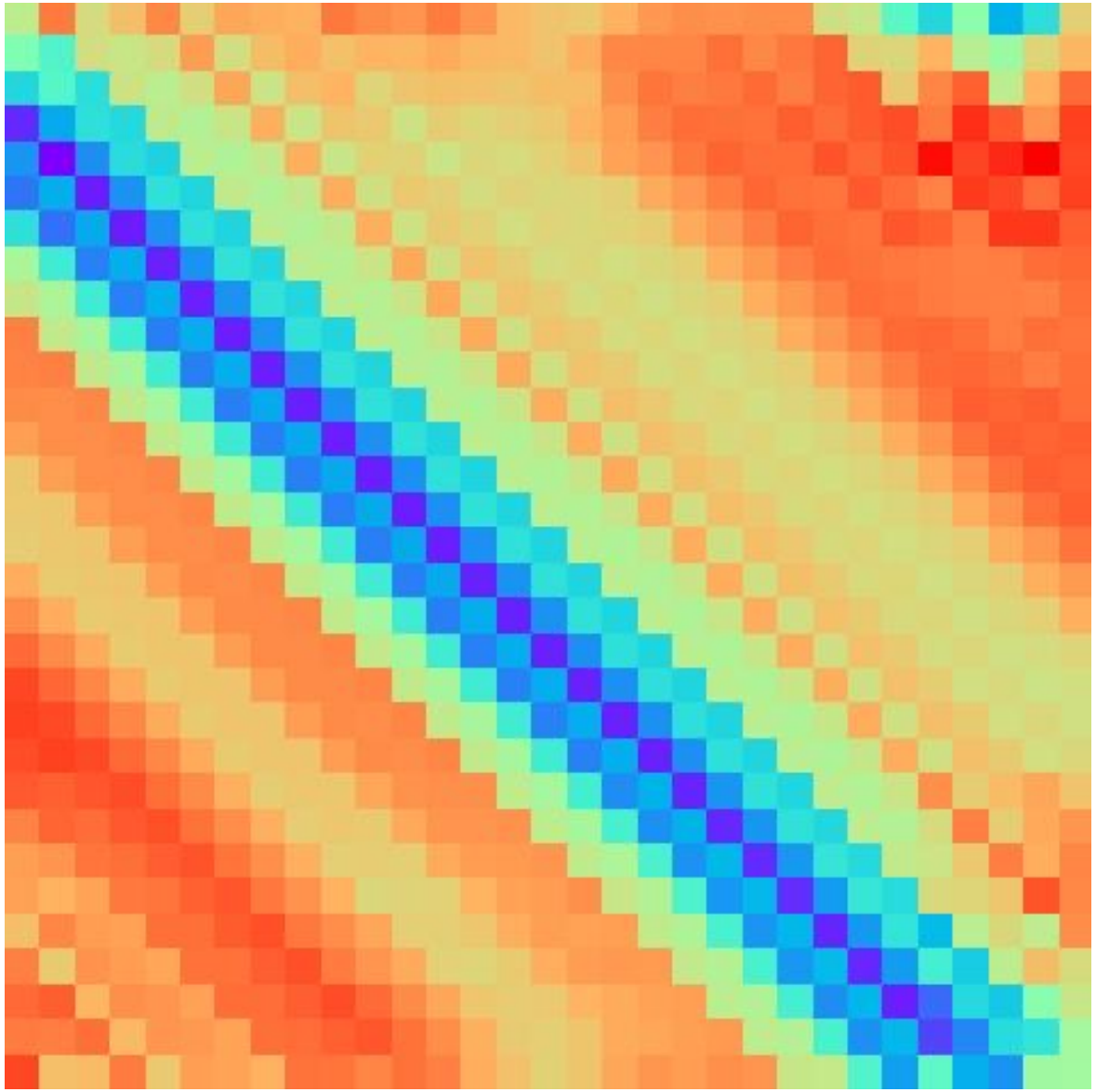




Reccurence Plot

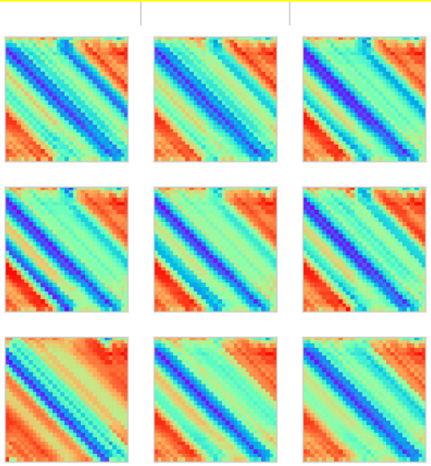


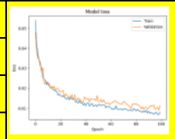
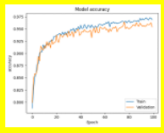
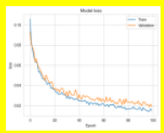
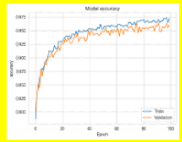
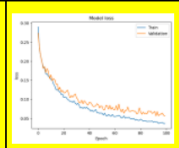
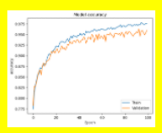
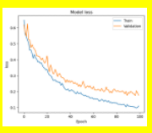
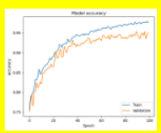
Accuracy (validation data)	Margin size			
	0.1	0.2	0.5	1
GASF	0.91719	0.90742	0.89531	0.87852
MTF	0.72383	0.71953	0.7168	0.68867
RP	0.92227	0.90078	0.89141	0.88398



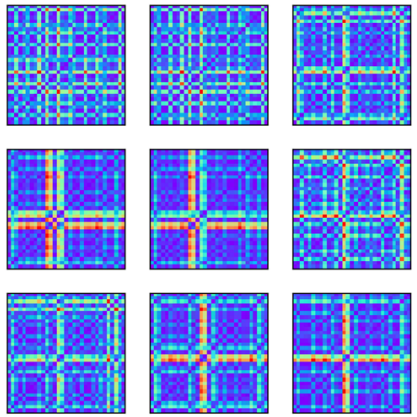
Pasiūlytas metodas konvertuoti skaitinius duomenis į vaizdus

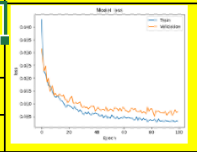
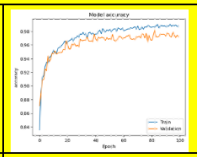
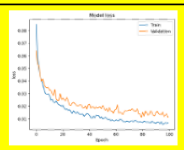
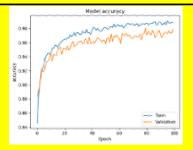
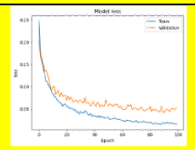
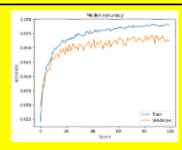
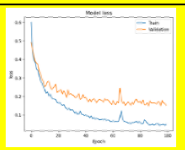
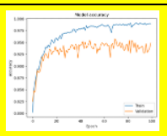
GAIT triplet



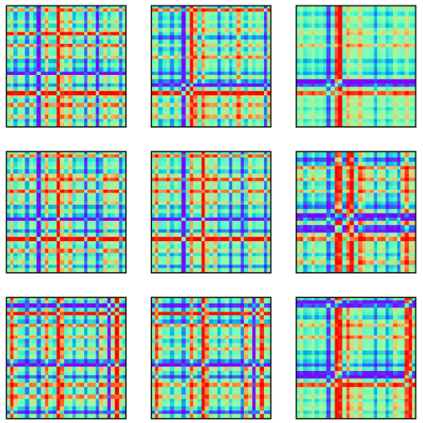
	Total time: 6309		Total time: 6289		Total time: 1757		Total time: 1767	
	training_0.1_margin.10epo.log		training_0.2_margin.10epo.log		training_0.5_margin.10epo.log		training_1_margin.10epo.log	
	AB_m0.1_e100.h5		AB_m0.2_e100.h6		AB_m0.5_e100.h7		AB_m1_e100.h8	
	Margin=0.1		Margin=0.2		Margin=0.5		Margin=1	
								
	Testavimui	Validavimui	Testavimui	Validavimui	Testavimui	Validavimui	Testavimui	Validavimui
Accuracy	0.89365	0.96172	0.8812	0.96641	0.8697	0.96484	0.8226	0.96719
AP_ED	0.16366	0.15667	0.38271	0.32062	0.66785	0.53447	0.79322	0.48577
AN_ED	0.41984	0.58718	0.86791	1.11733	1.38265	1.63223	1.4721	1.7588
AP_STD	0.11327	0.09809	0.25176	0.19778	0.41964	0.35884	0.50698	0.4039
AN_STD	0.17847	0.2511	0.34429	0.37271	0.4493	0.37325	0.52037	0.37752
AN_CS	0.85412	0.81404	0.68735	0.64051	0.48737	0.45833	0.43367	0.43886
AP_CS	0.91817	0.92167	0.80865	0.83969	0.66608	0.73276	0.60339	0.75712

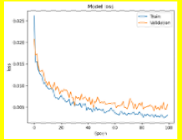
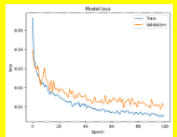
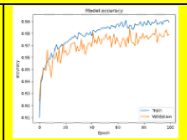
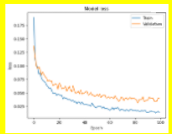
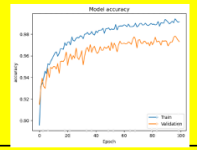
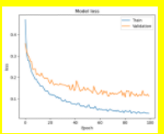
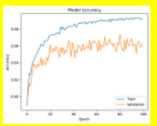
RP triplet



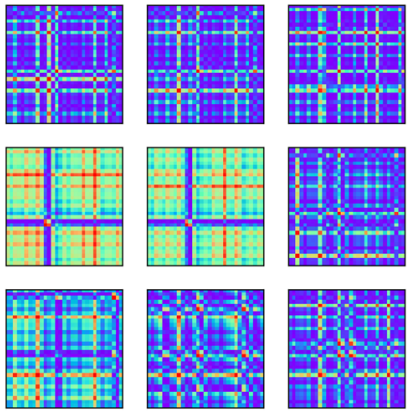
	Total time: 1947		Total time: 1917		Total time: 1955		Total time: 1953	
	RP_training_0.1_margin.10epo.log		RP_training_0.2_margin.10epo.log		RP_training_0.5_margin.10epo.log		RP_raining_1_margin.10epo.log	
	RP_m0.1_e100.h5		RP_m0.2_e100.h6		RP_m0.5_e100.h7		RP_m1_e100.h8	
	Margin=0.1		Margin=0.2		Margin=0.5		Margin=1	
								
	Testavimui	Validavimui	Testavimui	Validavimui	Testavimui	Validavimui	Testavimui	Validavimui
Accuracy	0.8956	0.9793	0.86815	0.97969	0.8587	0.98242	0.80245	0.97852
AP_ED	0.22202	0.17557	0.41205	0.33097	0.74671	0.47306	0.81296	0.3594
AN_ED	0.49735	0.68241	0.82158	1.13431	1.43524	1.72146	1.47756	1.8269
AP_STD	0.11698	0.08642	0.23675	0.17855	0.41419	0.30873	0.56749	0.36024
AN_STD	0.20211	0.24539	0.30908	0.31809	0.47042	0.34158	0.5522	0.34234
AN_CS	0.82016	0.78551	0.69159	0.63368	0.45451	0.45137	0.42737	0.45343
AP_CS	0.88899	0.91222	0.79397	0.83451	0.62665	0.76347	0.59352	0.8203

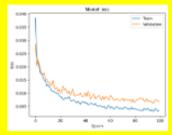
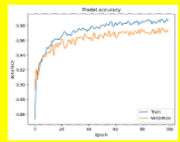
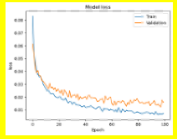
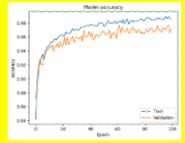
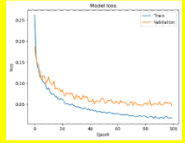
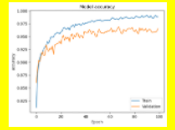
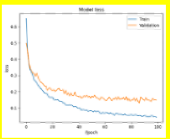
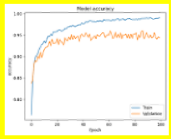
GADF triplet



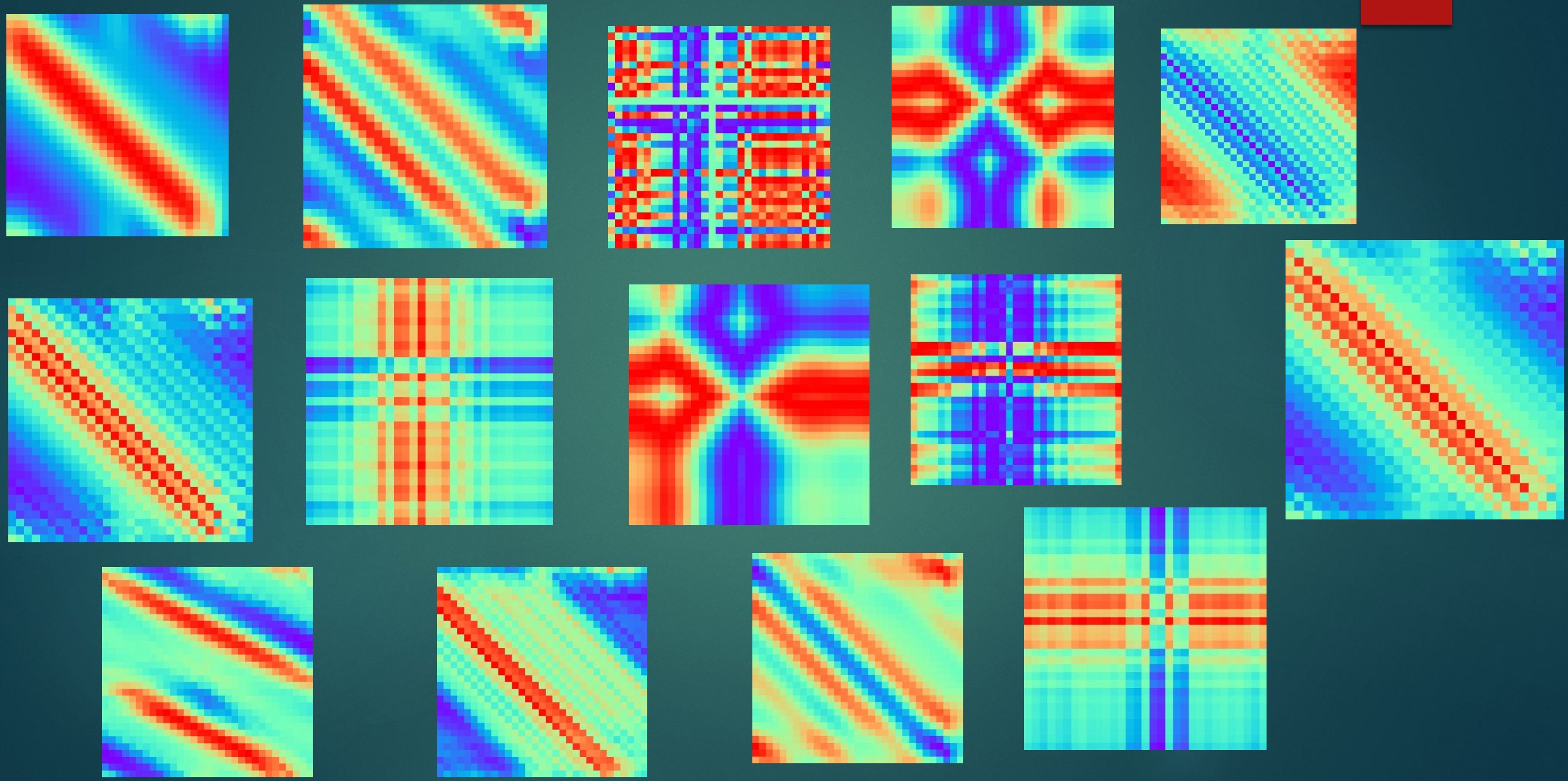
	Total time: 1977		Total time: 1856		Total time: 1768		Total time: 1967	
	GADF_training_0.1_margin.10epo.log		GADF_training_0.2_margin.10epo.log		GADF_training_0.5_margin.10epo.log		GADF_raining_1_margin.10epo.log	
	GADF_m0.1_e100.h5		GADF_m0.2_e100.h6		GADF_m0.5_e100.h7		GADF_m1_e100.h8	
	Margin=0.1		Margin=0.2		Margin=0.5		Margin=1	
								
	Testavimui	Validavimui	Testavimui	Validavimui	Testavimui	Validavimui	Testavimui	Validavimui
Accuracy	0.8991	0.98438	0.89915	0.98398	0.8675	0.99023	0.8169	0.98984
AP_ED	0.20823	0.19008	0.38448	0.34021	0.66125	0.45434	0.99435	0.32286
AN_ED	0.48002	0.72955	0.85966	1.20782	1.32183	1.7273	1.6887	1.87635
AP_STD	0.10739	0.09445	0.21645	0.17871	0.39774	0.30712	0.59024	0.30201
AN_STD	0.19379	0.2363	0.31945	0.3271	0.45245	0.323	0.45684	0.32365
AN_CS	0.82794	0.77009	0.68896	0.61299	0.50423	0.45459	0.32752	0.4502
AP_CS	0.89589	0.90496	0.80776	0.8299	0.66937	0.77283	0.50132	0.83857

GASF triplet

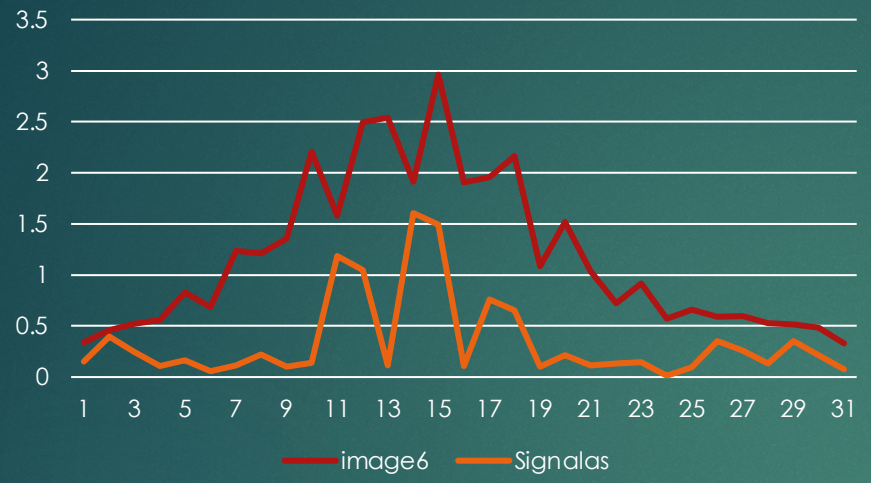


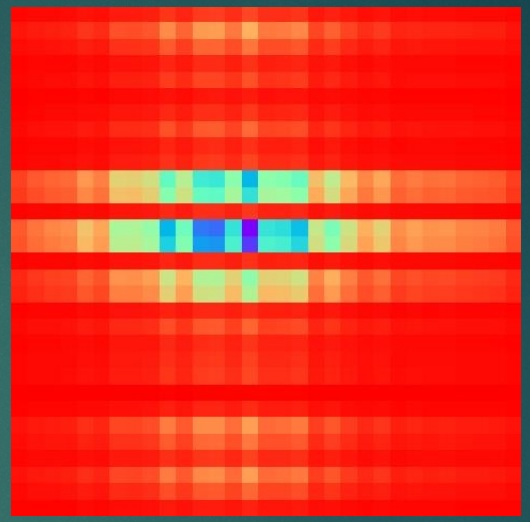
	Total time: 1979		Total time: 1891		Total time: 1955		Total time: 1991	
	GASF_training_0.1_margin.10epo.log		GASF_training_0.2_margin.10epo.log		GASF_training_0.5_margin.10epo.log		GASF_raining_1_margin.10epo.log	
	GASF_m0.1_e100.h5		GASF_m0.2_e100.h6		GASF_m0.5_e100.h7		GASF_m1_e100.h8	
	Margin=0.1		Margin=0.2		Margin=0.5		Margin=1	
								
	Testavimui	Validavimui	Testavimui	Validavimui	Testavimui	Validavimui	Testavimui	Validavimui
Accuracy	0.872	0.9875	0.871	0.98242	0.8386	0.98242	0.7927	0.99102
AP_ED	0.20455	0.16577	0.47131	0.36473	0.77677	0.49549	0.95292	0.34658
AN_ED	0.46292	0.66753	0.97841	1.27943	1.41581	1.72165	1.57008	1.86914
AP_STD	0.11806	0.08106	0.25468	0.19791	0.45604	0.30581	0.57386	0.34576
AN_STD	0.19271	0.23343	0.36656	0.34638	0.47405	0.33562	0.51754	0.31154
AN_CS	0.83313	0.79167	0.63757	0.58896	0.45185	0.44571	0.36925	0.44607
AP_CS	0.89772	0.91712	0.76434	0.81763	0.61161	0.75225	0.52354	0.82671

Pasiūlyto metodo modifikacijos



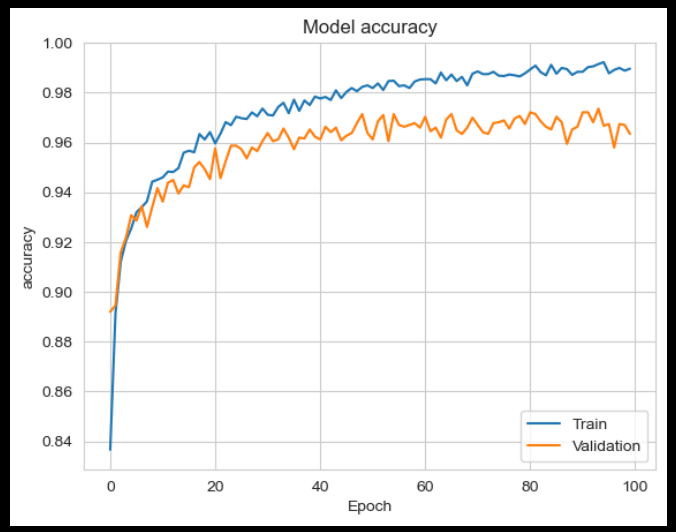
Pasiūlyto metodo modifikacija versija n

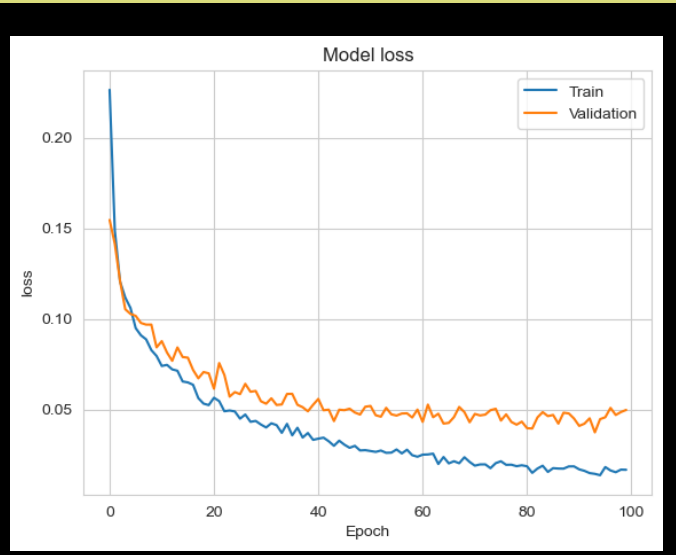




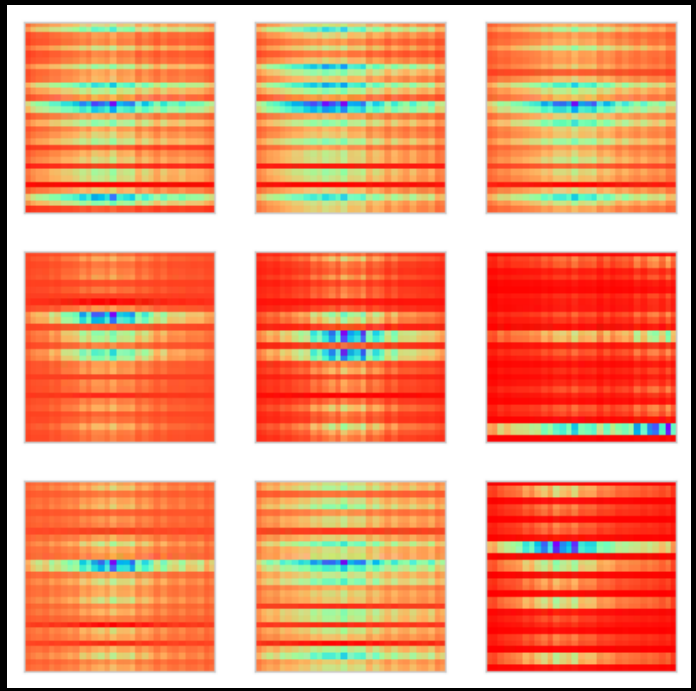
signalas = np.array([0.1491, 0.3979, 0.2488, 0.1069, 0.1674, 0.0605, 0.1169, 0.2212, 0.1043, 0.1417, 1.1885, 1.0468, 0.1146, 1.6055, 1.4909, 0.1067, 0.759, 0.6523, 0.1016, 0.2136, 0.112, 0.1349, 0.1484, 0.0135, 0.0932, 0.3515, 0.2583, 0.1338, 0.3509, 0.2171, 0.0742])

image6 = np.array([0.34105479148408757, 0.460597433714109 44, 0.5233945773583324, 0.5577215968832916, 0.8291349143264655, 0.682035814237058, 1.2357875878112694, 1.2117993876050017, 1.354308828828763, 2.2086671752035887, 1.5790510496578 71, 2.4952833156164753, 2.5422416228512006, 1.9150171176831916, 2.960993168828211, 1.9062349805844176, 1.955701169687063, 2.163594648684 33, 1.0836083633741564, 1.5158326216640519, 1.0377995064661623, 0.722507338590761, 0.9156124007895252, 0.5732721456792367, 0.657088484795757 7, 0.5886282796771238, 0.5994131218280678, 0.5262492114945743, 0.514198848299743 3, 0.4868423133649406 6, 0.3302565659517387])





	Total time: 1955	
	GAIT9_m0.5_best_e100.hdf5	
	Margin=0.5	
	Testavimui	Validavimui
Accuracy	0.866	0.98555
AP_ED	0.84067	0.50124
AN_ED	1.48083	1.758
AP_STD	0.47572	0.32396
AN_STD	0.43385	0.30092
AN_CS	0.41962	0.43519
AP_CS	0.57966	0.74938
EER	0.229	0.04883
AUC	0.83289	0.98481



Kito pusmečio darbo planas

- Publikacija mokslo leidinyje, turinčiame cituojamumo rodiklį Clarivate Analytics Web of Science duomenų bazėje (planuojama iki 2023 m. rugsėjo mėn.);
- Dalyvauti "25TH INTERNATIONAL CONFERENCE ON HUMAN-COMPUTER INTERACTION" konferencijoje, kuri vyks Danijoje (Kopenhaga) 2023 m. liepos 23-28 dienomis (ACCEPTED SUBMISSIONS WILL BE INCLUDED IN THE CONFERENCE PROCEEDINGS to be published by Springer in the Lecture Notes in Computer Science (LNCS) or Lecture Notes in Artificial Intelligence (LNAI) series);
- Dalyvauti "18TH IBERIAN CONFERENCE ON INFORMATION SYSTEMS AND TECHNOLOGIES (CISTI) " konferencijoje, kuri vyks Portugalijoje (Aveira) 2023, pp. 1-6. IEEE. (Submitted, under review).
- Metodologijos bei sudaryto metodo pritaikymas praktinių uždavinių sprendimui;
- Gautų duomenų analizė, rezultatų apibendrinimas, išvadų parengimas.

Jei neužduosi teisingų klausimų, negausi teisingų atsakymų. Teisingai užduotame klausime dažnai jau slypi atsakymas.

EDWARD HODNET

arnoldas.budzys@mif.stud.vu.lt