Vilniaus universitetas
Duomenų mokslo ir skaitmeninių technologijų institutas

Informatikos krypties doktorantų atestacinė konferencija
Veiklos ataskaita už 2022 m. kovo 24 d. – 2022 m. rugsėjo 30 d.

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

dokt. Arnoldas BUDŽYS – Informatika N 009
Studijų metai: II
Darbo vadovas: dr. Viktor Medvedev
Doktorantūros pradžios ir pabaigos metai: 2020–2024

2021–2022 m.

STUDIJŲ PLANAS IR JO VYKDYMO SUVESTINĖ

Studijų metai	Egzaminai [1]		Dalyvavimas konferencijose [2]		Publikacijos [3]		
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė [4]
I (2020/2021)	1	1					
II (2021/2022)	**3**	**3**	**1**	**1+1**	**1** (be cituojamumo rodiklio)	0	Procese
III (2022/2023)			**1**		1+1 (skola iš II metų)		
IV (2023/2024)			1		1		
Iš viso:	4	4	3	2	3		

2021–2022 m.

ATASKAITINIŲ METŲ DARBO PLANAS IR JO SUVESTINĖ

Egzaminai	
Planas	Įvykdyta
Fundamentalieji informatikos ir informatikos inžinerijos metodai	Išlaikytas
Mašininis mokymasis	Išlaikytas
Gilieji neuroniniai tinklai	Išlaikytas

2021–2022 m.

ATASKAITINIŲ METŲ DARBO PLANAS IR JO SUVESTINĖ

Dalyvavimas konferencijose		
Planas	Įvykdyta	Konferencijos tipas
Data Analysis Methods for Software Systems 2021 m. gruodžio 1–3 d., Druskininkai	User Behaviour Analysis Based on Similarity Measures to Detect Anomalies Data Analysis Methods for Software Systems 2021 m. gruodžio 1-3 d., Druskininkai Autoriai: Arnoldas Budžys, Viktor Medvedev, Olga Kurasova Įvertintas kaip geriausias posteris	Nacionalinė

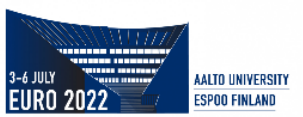
Publikacijos			
Planas	Įvykdyta	Būklė	Publikacijos tipas
	Budžys, Arnoldas; Medvedev, Viktor; Kurasova, Olga. User behaviour analysis based on similarity measures to detect anomalies // DAMSS: 12th conference on data analysis methods for software systems, Druskininkai, Lithuania, December 2–4, 2021. Vilnius : Vilnius University Press, 2021. ISBN 9786090706732. eISBN 9786090706749. p. 8. DOI: 10.15388/DAMSS.12.2021.	Publikuota	Be cituojamumo rodiklio

2021–2022 m.

ATASKAITINIŲ METŲ DARBO PLANAS IR JO SUVESTINĖ

Dalyvavimas konferencijose		
Planas	Įvykdyta	Konferencijos tipas
32nd European Conference on Operational Research (EURO **XXXII**)], Espoo, Finland, July 3-6, 2022	Deep learning-based prevention of insider threats using user behavioral keystroke biometrics Autoriai: Arnoldas Budžys, Olga Kurasova, Viktor Medvedev	Tarptautinė

Publikacijos			
Planas	Įvykdyta	Būklė	Publikacijos tipas
	Budžys, Arnoldas; Kurasova, Olga; Medvedev, Viktor. Deep learning-based prevention of insider threats using user behavioral keystroke biometrics // EURO 2022: [32nd European Conference on Operational Research (EURO XXXII)], Espoo, Finland, July 3-6, 2022 : abstract book. Espoo : Aalto university, 2022. ISBN 9789519525419. p. 144. Prieiga per internetą: https://www.euro-online.org/conf/admin/tmp/program-euro32.pdf.	Publikuota	Konferencijų medžiaga



Mr. Arnoldas Budžys
Institute of Data Science and Digital Technologies
Vilnius university
Lithuania

Espoo, September 26, 2022

Certificate of Attendance

This is to certify that

Mr. Arnoldas Budžys

attended the 32nd European Conference on Operational Research (EURO 2022), in Espoo, Finland from the 3rd to the 6th of July, 2022, presenting the following paper, co-authored with Prof. Olga Kurasova, Dr. Viktor Medvedev:

Deep Learning-Based Prevention of Insider Threats Using User Behavioral Keystroke Biometrics

Sincerely,

Antti Punkka
Conference Chair of the Organising Committee
of the 32nd European Conference
on Operational Research (EURO 2022)

2021–2022 m.

ATASKAITINIŲ METŲ DARBO PLANAS IR JO SUVESTINĖ

Tarptautinė vasaros/žiemos stovykla		
Planas	Įvykdyta	Stovyklos tipas
Eastern European Machine Learning Summer School 6-14 July 2022, Vilnius Lithuania	Application of Keystroke-Based Behavioral Biometrics for Insider Threat Prevention Using Machine Learning Techniques	Tarptautinė vasaros mokykla

CERTIFICATE OF PARTICIPATION

EEML CERTIFIES THAT

Arnoldas Budžys

has attended the sessions of EEML2022 summer school.

EEML

DOINA PRECUP
EEML Organizer

Karolina Dziugaite
EEML Organizer

VIORICA PATRAUCEAN
EEML Organizer

RAZVAN PASCANU
EEML Organizer

Linas Petkevičius
EEML Organizer

Dovydas Čeilutka
EEML Organizer

Jev Gamper
EEML Organizer

Mokslinių tyrimų ir disertacijos rengimo planas:

2020–2021 m.

Mokslinių tyrimų disertacijos tema apžvalga ir analizė (Lietuvoje ir užsienyje):

- **Disertacijos tyrimo objekto detalizavimas;**
- **Atlikti mašininio mokymosi metodų taikymo kompiuterių tinkluose analitinę apžvalgą;**
- **Nustatyti (identifikuoti) mokslines problemas, kylančias uždaviniuose, susijusiuose su įsilaužimų prevencija kompiuterių tinkluose taikant mašininio mokymosi metodus;**
- **Tyrimo tikslo suformavimas.**

Mokslinių tyrimų ir disertacijos rengimo planas:

2021–2022 m.

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. Įsilaužimų prevencijos atpažinimo mašininio mokymosi metodo sukūrimas ir/ar testavimas.**

Mokslinių tyrimų ir disertacijos rengimo planas:

2022–2023 m.

- **2.3. Empirinis tyrimas:**
- 2.3.1. Sudarytų metodų pritaikymas praktinių uždavinių sprendimui.
- 2.3.2. Gautų duomenų analizė, rezultatų apibendrinimas, išvadų parengimas.

Mokslinių tyrimų ir disertacijos rengimo planas:

2023–2024 m.

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.

Mokslinių tyrimų ir disertacijos rengimo planas:

2024 m. birželio mėnesį

- Daktaro disertacijos parengimas ir svarstymas padalinyje

2024 m. rugsėjo mėnesį

- Daktaro disertacijos gynimas

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.

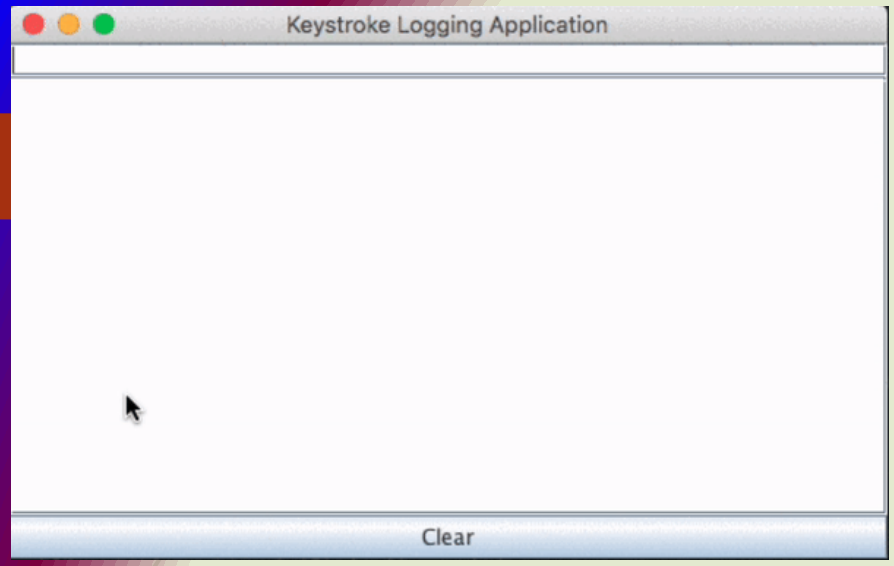
Literatūros analitinės apžvalgos apibendrinimas

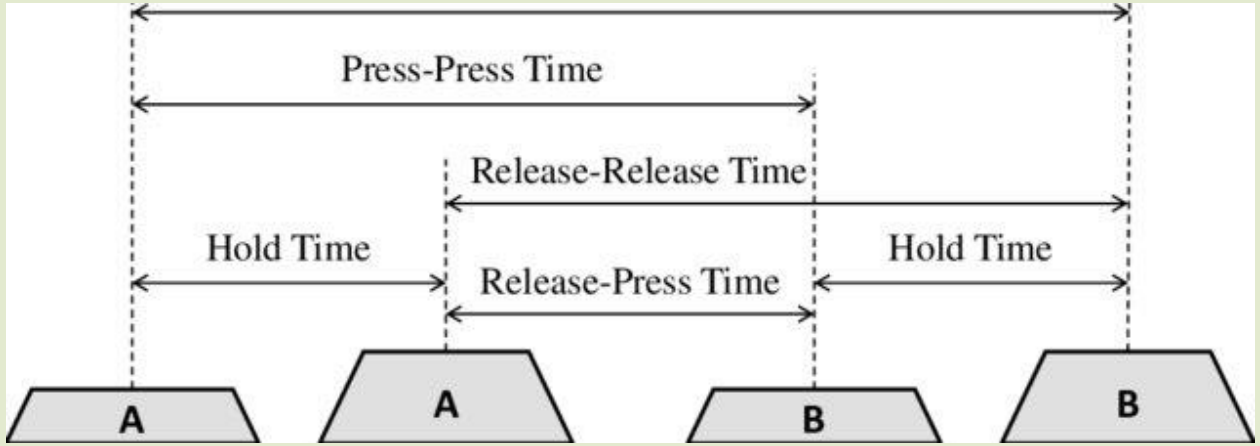
- Pastaraisiais dešimtmečiais buvo pasiūlyta keliolika mašininio mokymosi algoritmais, dirbtiniais neuroniniais tinklais grįsti metodai sistemos vartotojo autentifikavimui pagal jo klaviatūros sudarytus biometrinius elgsenos duomenis.
- Biometrinių duomenų rinkimas panaudojant klavišų paspaudimus, turi didelį potencialą nustatant asmens tapatybę. Unikalų klavišų paspaudimo profilį galima sukonstruoti iš įvairių spausdinimo ypatybių, pvz., spausdinimo greičio, trukmės tarp paspaustų klavišų ir spaudžiamų klavišų, naudojamų klavišų ir pan. Toks duomenų rinkimo metodas nereikalauja papildomų veiksnių atliekant veiksmus darbo stotyje ar bet kurioje kitoje platformoje.
- Vartotojo autentifikavimas naudojant klaviatūros biometrinius duomenis skirstomas į dvi kategorijas:
  - Pirminį (Statinį) autentifikavimą (SA)
  - Nepertraukiamą autentifikavimą (NA).

Atliktas teorinis tyrimas

- Tarpusavyje palygintos dirbtinių neuroninių tinklų (*Long short-term memory*, LSTM; *MultiLayer Perceptron*, MLP; *Convolutional Neural Network*, CNN; MLP + *KerasTuner* (KT) optimizatorius) grįstos metodikos vartotojų klasifikavimui atlikti.
  - Eksperimentų metu buvo keičiami dirbtinių neuroninių tinklų parametrai bei siekiama padidintį CMU duomenų aibės vartotojų klasifikavimą.
- Pasiūlytas metodas, kuriame duomenys transformuojami į vaizdinį pavidalą, panaudojant Siamo neuroninius tinklus (angl. *Siamese Neural Network*, *SNN*). Gauti eksperimentinio tyrimo rezultatai parodo, kad galima pagerinti vartotojų klasifikavimo rezultatą, lyginant su mašininio mokymosi algoritmais, bei klasikiniais dirbtiniais neuroniniais tinklais kuomet klasifikavimui naudojami skaitiniai duomenys.

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).

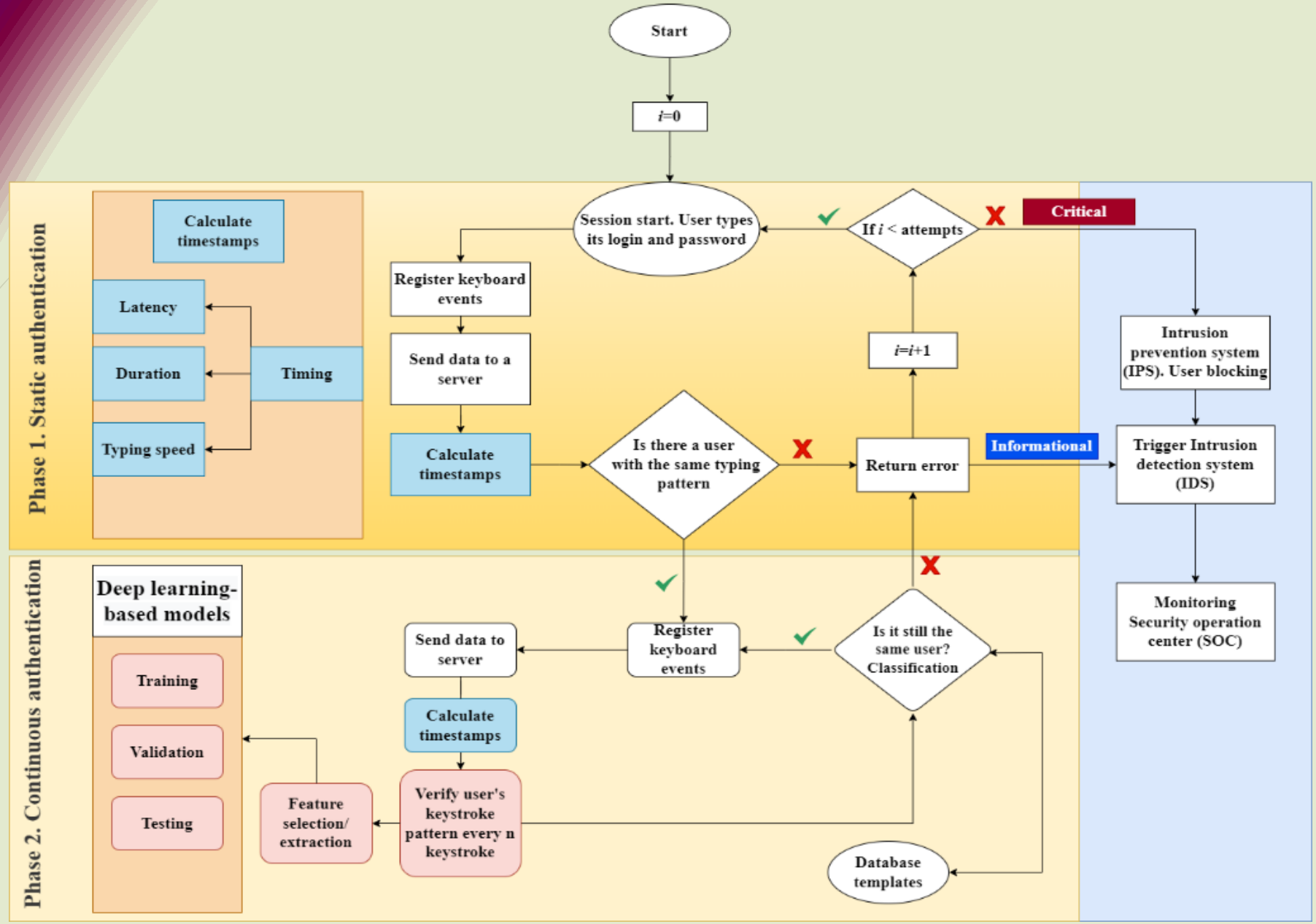
Duomenų aibės

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

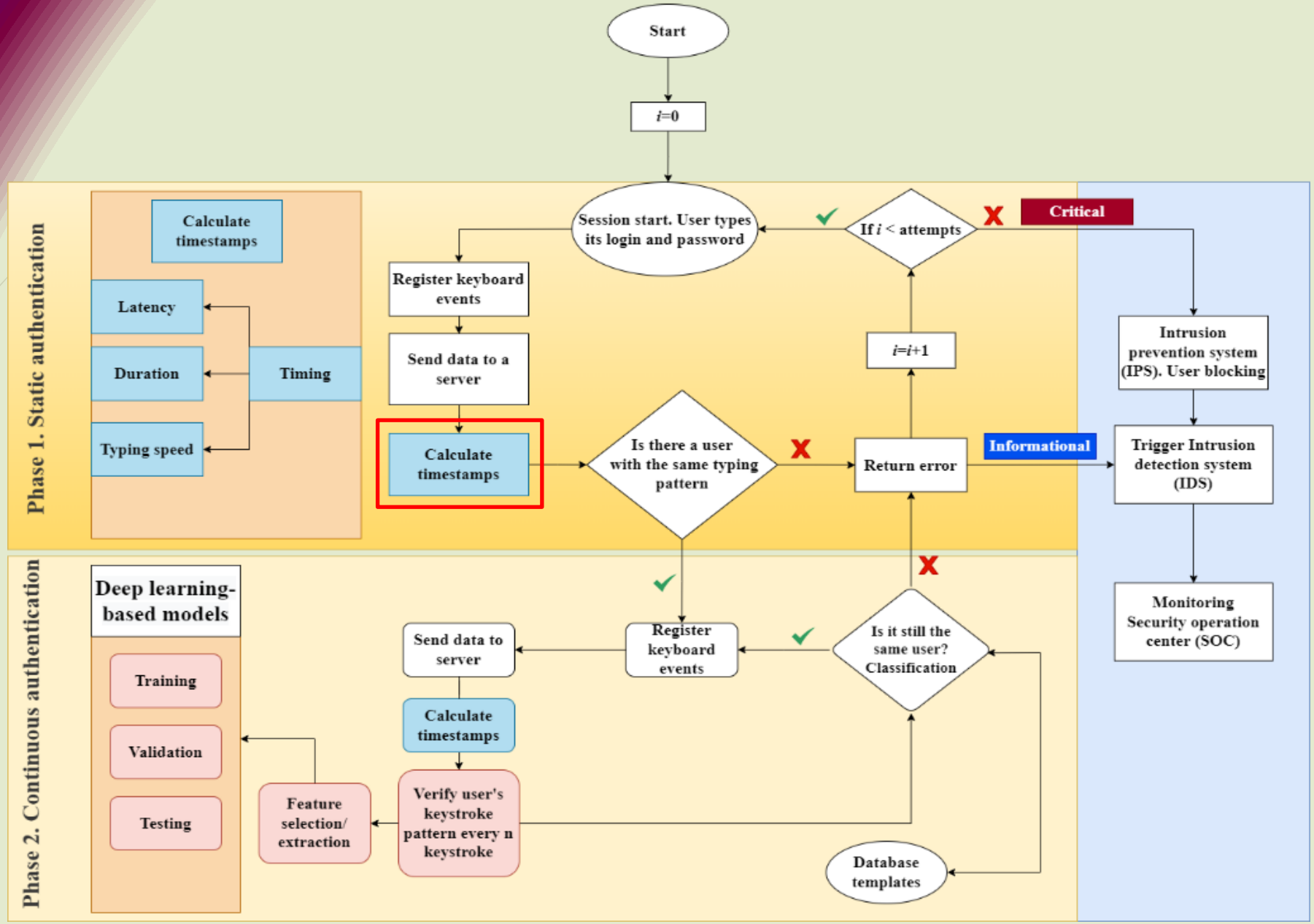
References: **2.** Killourhy, K. S., & Maxion, R. A. (2009, June). Comparing anomaly-detection algorithms for keystroke dynamics. In *2009 IEEE/IFIP International Conference on Dependable Systems & Networks* (pp. 125-134). IEEE. **3.** Giot, R., El-Abed, M., & Rosenberger, C. (2009, September). Greyc 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

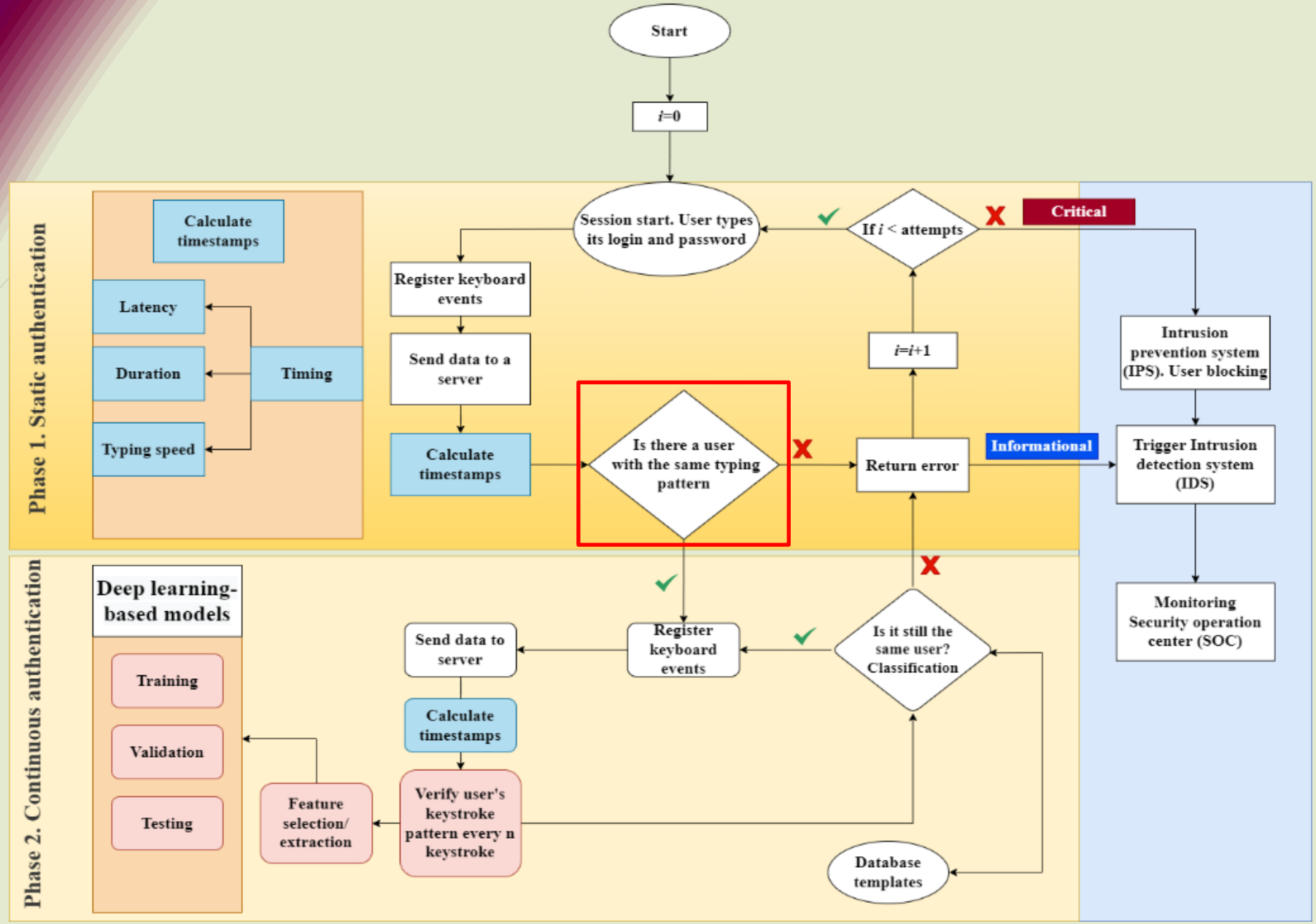
- Sistemos vartotojui autentifikuojantis sistemoje yra tikrinamas jo įvesto slaptažodžio laiko žymos ir lyginamos su jau duomenų bazėje esančiu vartotojo šablonu.
- Prisijungus prie sistemos, vartotojo klaviatūros biometrika yra toliau stebima ir kaskart tikrinama ar pirminiame etape autentifikavęsis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



2 pav. Vartotojo pirminio (statinio) ir nepertraukiamo autentifikavimo schema



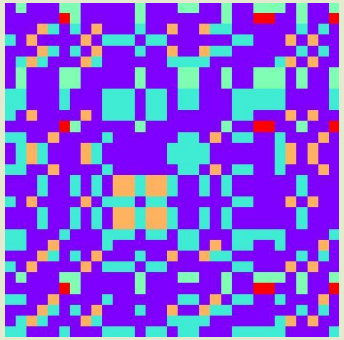
2 pav. Vartotojo pirminio (statinio) ir nepertraukiamo autentifikavimo schema



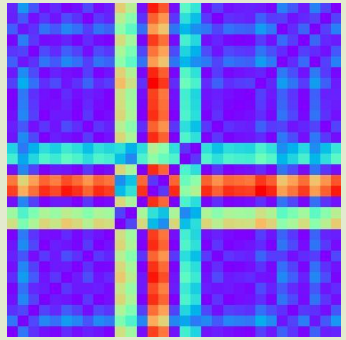
2 pav. Vartotojo pirminio (statinio) ir nepertraukiamo autentifikavimo schema

No Image to Image

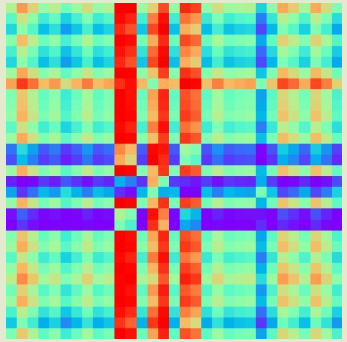
1. Markov Transition Field (MTF)[6]



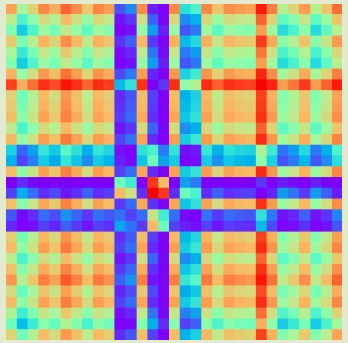
2. Reccurence Plots (RPs)[7]



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



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

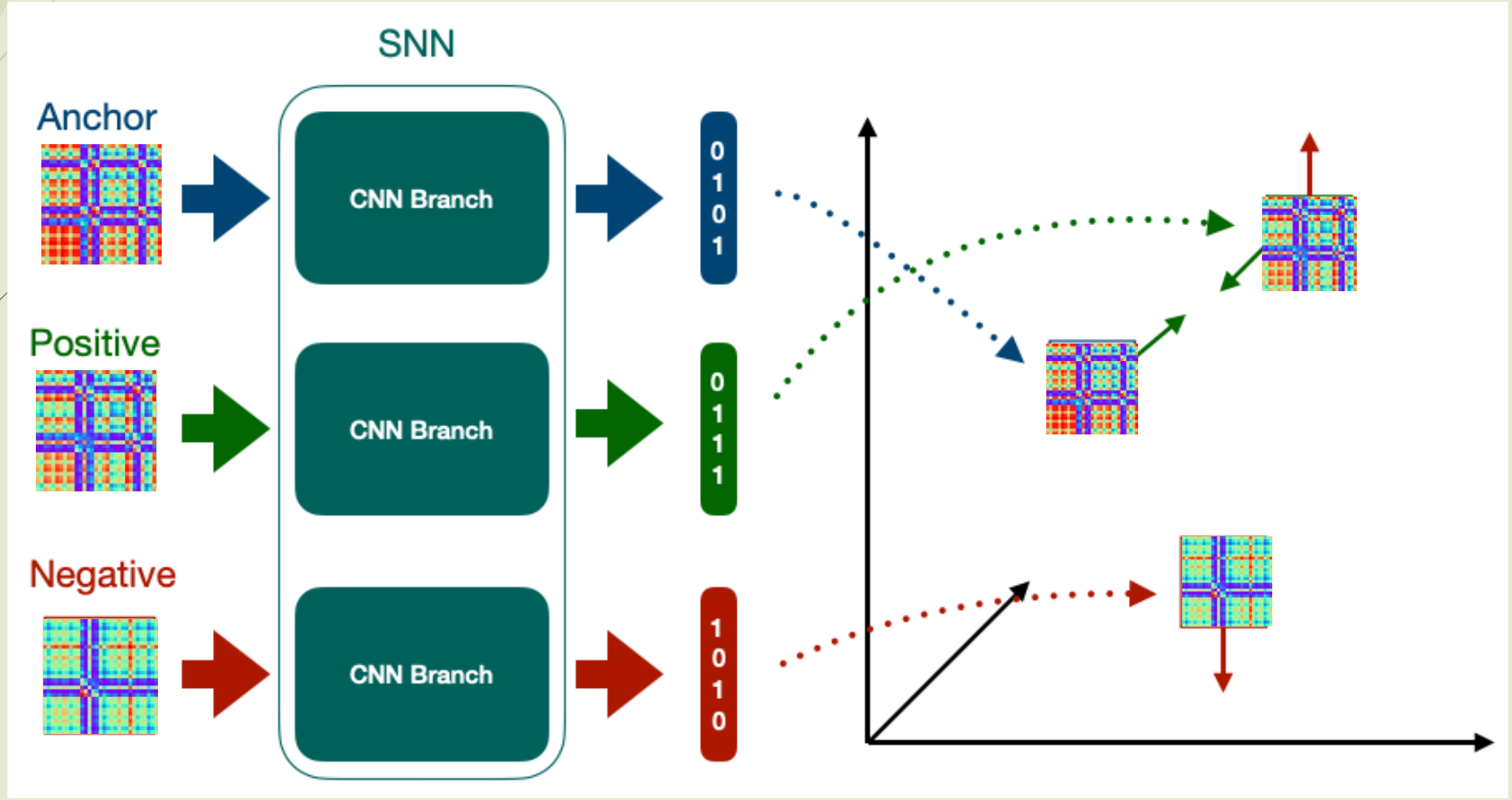


5. Vector to image (Vec2im)[9]

6. Image fusion (combined)

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.

Siamo neuroniniai tinklai (angl. *Siamese Neural Network*, SNN)



3 pav. Siamo tinklo metodika[10]

References: **10**. How to train your Siamese Neural Network, https://towardsdatascience.com/how-to-train-your-siamese-neural-network-4c6da3259463

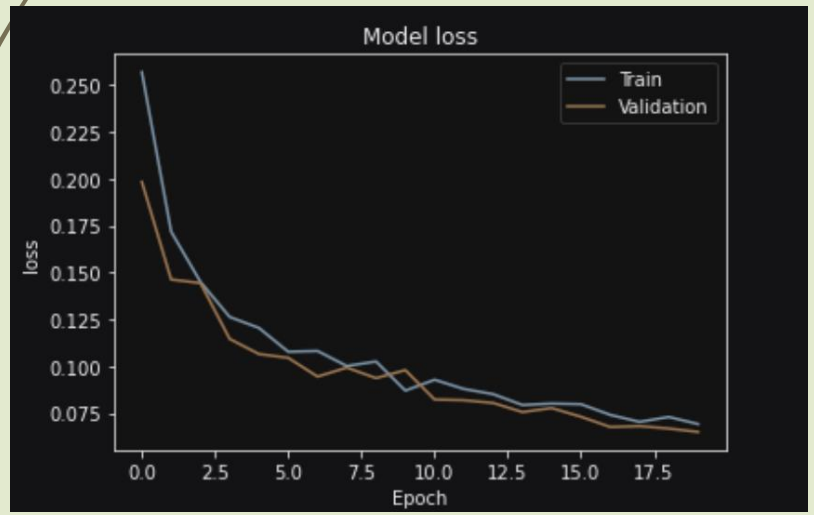
Duomenų paruošimas (CMU)

80 % tinklo apmokymui
20 % tinklo validavimui

10 % testavimui

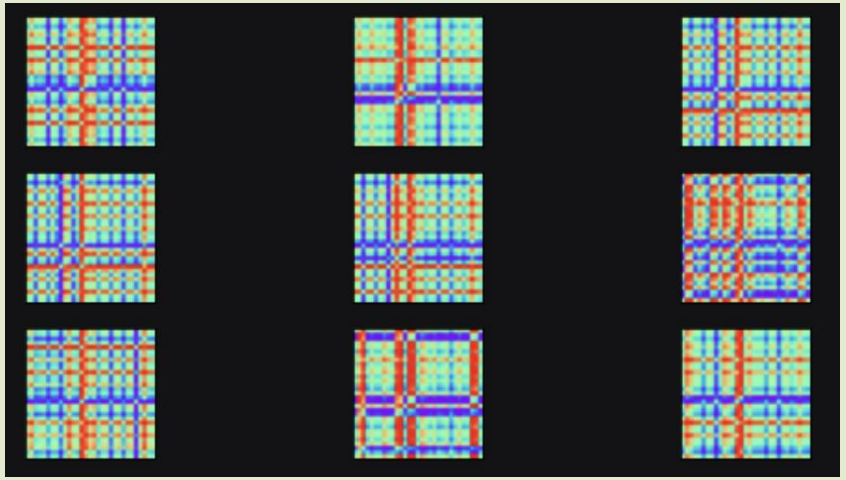
Tinklo mokymas ir validavimas

```
history = siamese_model.fit(train_dataset,
epochs=20,
validation_data=val_train_dataset)
```



4 pav. Siamo tinklų apmokymas

Tinklo testavimas



5 pav. Validavimo duomenys: Anchor, positive, negative

Results:

Positive similarity: 0.99479556
Negative similarity 0.99377453

Kito pusmečio darbo planas

- Apžvalginio straipsnio rašymas (periodiniame recenzuojame mokslo žurnale arba konferencijų medžiagoje) (Procese);
- 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 (*Abstract deadline*: 2022.10.21, *Proceedings deadline*: 2023.02.03. 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).

Kito pusmečio darbo planas

- Mašininio mokymosi metodų, naudojamų kompiuterių tinkluose įsilaužimų prevencijai, tyrimas;
- Įsilaužimų prevencijos atpažinimo mašininio mokymosi metodo sukūrimas ir/ar testavimas;
- Sudarytų metodų pritaikymas praktinių uždavinių sprendimui;
- Gautų duomenų analizė, rezultatų apibendrinimas, išvadų parengimas.

KLAUSIMAI?

arnoldas.budzys@mif.stud.vu.lt