Vilniaus universitetas
Duomenų mokslo ir skaitmeninių technologijų institutas

INFORMATIKOS KRYPTIES DOKTORANTŲ ATESTACINĖ KONFERENCIJA
VEIKLOS ATASKAITA UŽ 2023 M. KOVO 21 D. – 2023 M. RUGSĖJO 27 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**	**0**	**neįteikta****	**1**	**1+1***	Publikuota
IV (2023/2024)					1					
Iš viso:	2	3	1	2	2	0		1	2	

*HCI2023 – publikuota (CA WoS, Springer), CISTI2023 – publikuota (CA WoS, IEEE)

** Publikacija parengta (pridėta prie mokslinės ataskaitos kaip 3 priedas), vyksta korektūros, planuojama įteikti į IEEE Access, WoS IF iki 2023 m. rugsėjo 30 d.

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

Dalyvavimas konferencijose 2022/2023 (II pusmetis)		
Planas	Įvykdyta	Konferencijos tipas
25th Human Computer Interacting international conference Kopenhaga, Danijos karalystė 2023 m. liepos 23-28 d.	**Behavioral biometrics authentication in critical infrastructure using siamese neural networks** 25th Human Computer Interacting international conference 2023 m. liepos 23-28 d., Danijos Karalystė Autoriai: Arnoldas Budžys, Olga Kurasova, Viktor Medvedev	Tarptautinė

Publikacijos 2022/2023 (II pusmetis)			
Planas	Įvykdyta	Būklė	Publikacijos tipas
18th Iberian Conference on Information Systems and Technologies 2023 m. birželio 20–23 d. Aveiro, Portugalija	**Medvedev, Viktor; Budžys, Arnoldas; Kurasova, Olga**. Enhancing keystroke biometric authentication using deep learning techniques // 2023 18th Iberian Conference on Information Systems and Technologies (CISTI), 20-23 June, Aveiro, Portugal, 2023 : proceedings. New York : IEEE, 2023. ISBN 9798350305272. eISBN 9789893347928. ISSN 2166-0727. eISSN 2166-0727. p. [1-6]. DOI: 10.23919/CISTI58278.2023.10211344.	Publikuota	CA WoS duomenų bazėje be cituojamumo rodiklio

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

Publikacijos 2022/2023 (II pusmetis)			
Planas	Įvykdyta	Būklė	Publikacijos tipas
25th Human Computer Interacting international conference Kopenhaga, Danijos karalystė 2023 m. liepos 23-28 d.	**Budžys, Arnoldas; Kurasova, Olga; Medvedev, Viktor**. Behavioral biometrics authentication in critical infrastructure using siamese neural networks // HCI for cybersecurity, privacy and trust: 5th international conference, HCI-CPT 2023, held as part of the 25th HCI international conference, HCII 2023. Copenhagen, Denmark, July 23–28, 2023 : proceedings. Cham : Springer, 2023. ISBN 9783031358210. eISBN 9783031358227. p. 309-322. (Lecture notes in computer science, ISSN 0302-9743, eISSN 1611-3349 ; vol. 14045). DOI: 10.1007/978-3-031-35822-7_21.	Publikuota	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.**	Pateiktas naujas autoriaus pasiūlytas metodas (GAFMAT) skaitiniams duomenims konvertuoti į vaizdinius. Siekiant įvertinti šio naujo metodo veiksmingumą, atlikta išsami lyginamoji analizė naudojant Siamo neuroninius tinklus, gauti rezultatai palyginti su esamomis metodikomis, aprašytomis literatūroje. 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.

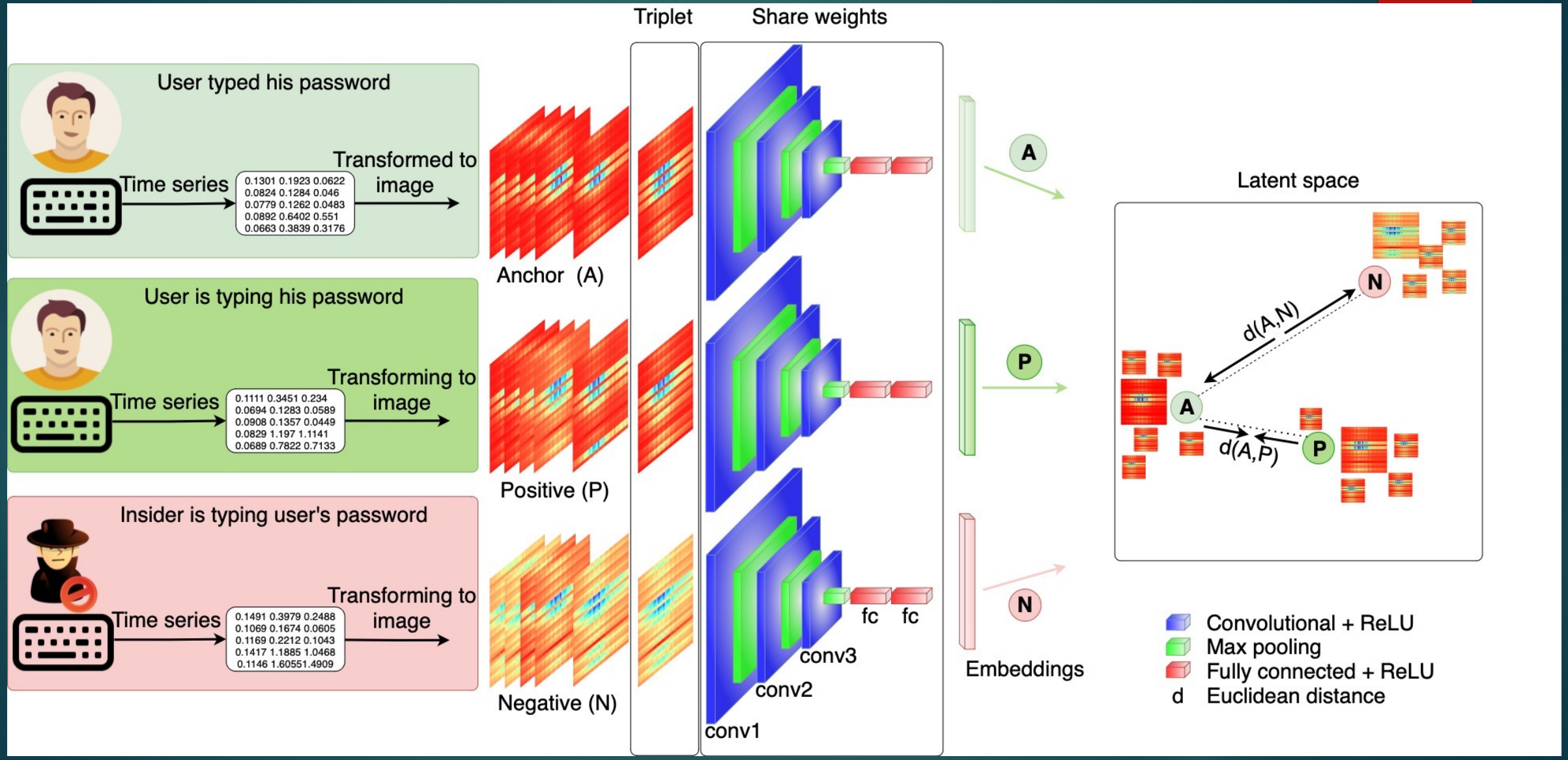
Tikslas:

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

Tyrimo uždaviniai

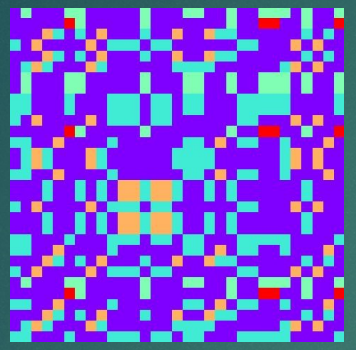
- Atlikti išsamią literatūros analitinę apžvalgą, siekiant identifikuoti tinkamus metodus anomalinių įvykių identifikavimui ir insaiderio užkardymui kompiuterių tinkluose;
- Atlikti skirtingų mašininio mokymosi metodų, skirtų anomalinių įvykių identifikavimui ir insaiderio užkardymui kompiuterių tinkluose, 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.

Metodologija

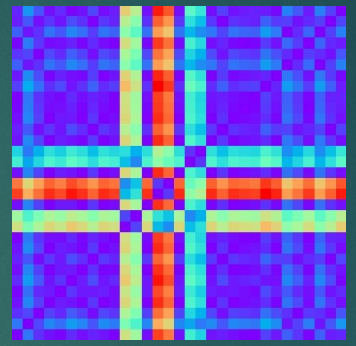


No Image to Image

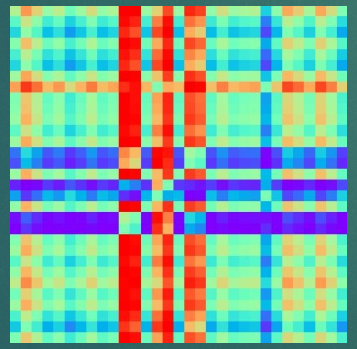
1. Markov Transition Field (MTF)[6]



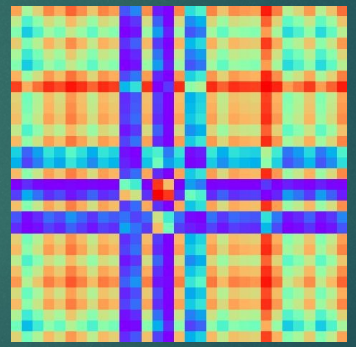
2. Reccurence Plots (RPs)[7]



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.

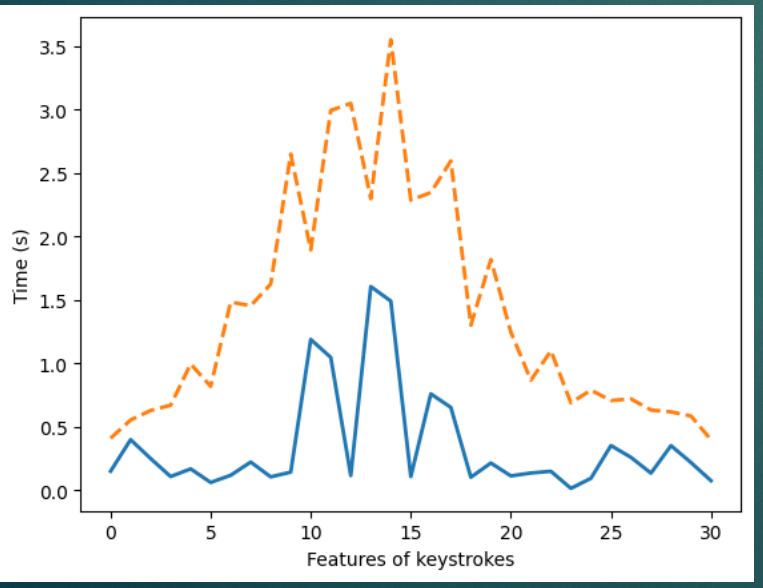
GAFMAT

Algorithm 1 Gabor Filter algorithm

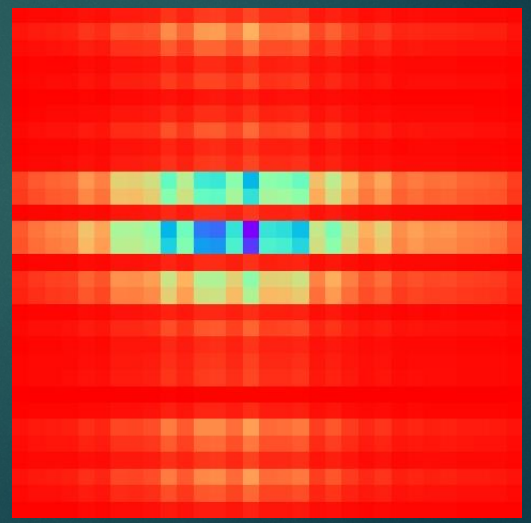
function GaborFilter($discrete\_signal, \sigma, \theta, \lambda, \psi, \gamma$)
  $n \leftarrow$ length of *discrete_signal*
  Initialize x as array of size n generating evenly-spaced values in an interval $(-3\sigma, 3\sigma)$
  $x \leftarrow x \cdot \cos(\theta)$
  Initialize *gabor* as an empty array of size n
  $gabor \leftarrow \exp\left(-0.5 \cdot \left(\frac{x}{\sigma}\right)^2\right) \cdot \cos\left(2\pi \cdot \frac{x}{\lambda} + \psi\right)$
  $gabor \leftarrow \frac{gabor}{\sqrt{\sum_{i=0}^{n-1} gabor[i]^2}}$
  $gabor \leftarrow Convolution(discrete\_signal, gabor)$
return *gabor*
end function

Algorithm 2 GAFMAT algorithm

Require: $discrete\_signal, \sigma\_list, \theta\_list, \lambda\_list, \psi\_list, \gamma\_list$
$n \leftarrow$ length of *discrete_signal*
image $\leftarrow$ create zero array of size n
combinations $\leftarrow$ *CartesianProduct*($\sigma\_list, \theta\_list, \lambda\_list, \psi\_list, \gamma\_list$)
▷ The set of all possible pairs (see Table 2)
for each $(\sigma, \theta, \lambda, \psi, \gamma)$ in *combinations* **do**
  gabortemp $\leftarrow$ *gabor*($discrete\_signal, \sigma, \theta, \lambda, \psi, \gamma$)
▷ see Algorithm 1
  gabor $\leftarrow$ *transpose*(*gabortemp*)
  image2D $\leftarrow$ *OuterProduct*(*image*, *gabor*)
end for
return *image2D*

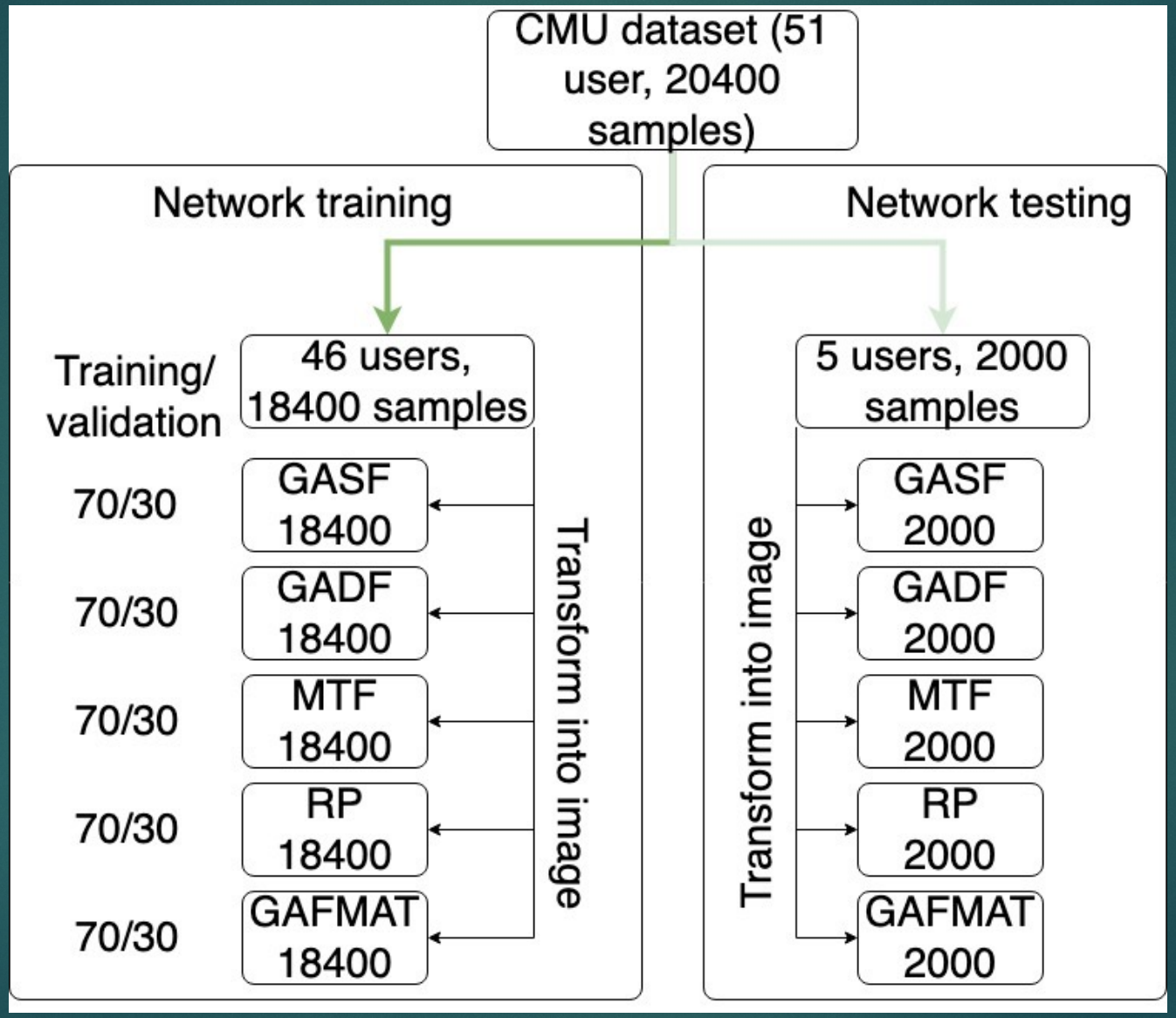


$$image2D = \begin{bmatrix} a_1b_1 & a_1b_2 & \cdots & a_1b_n \\ a_2b_1 & a_2b_2 & \cdots & a_2b_n \\ \vdots & \vdots & \ddots & \vdots \\ a_nb_1 & a_nb_2 & \cdots & a_nb_n \end{bmatrix}$$



GABOR FILTER MATRIX TRANSFORMATION

Data Preparation



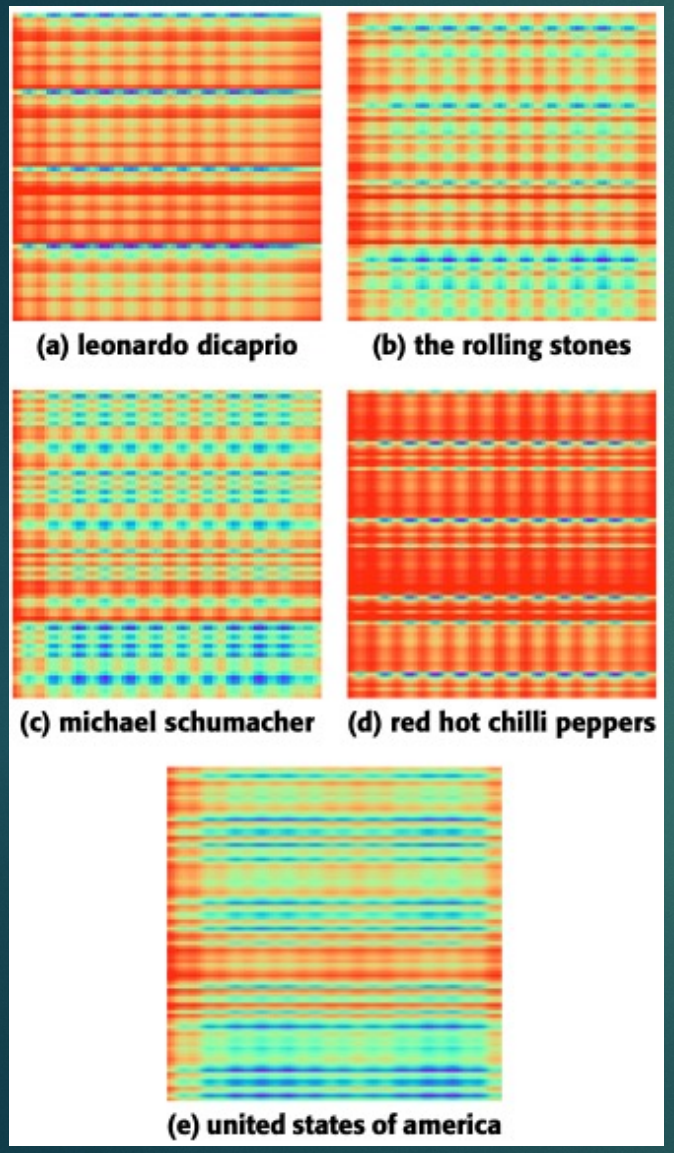
Results

	GADF		GASF		RP		MTF		GAFMAT	
	Testavimui	Validavimui	Testavimui	Validavimui	Testavimui	Validavimui	Testavimui	Validavimui	Testavimui	Validavimui
Accuracy	0.868	**0.99077**	0.854	0.98473	0.829	**0.98331**	0.854	0.94744	0.866	**0.98935**
AP_ED	0.73164	0.44127	0.86555	0.47255	0.84481	0.43633	0.86555	0.56487	0.83616	0.486
AN_ED	1.41323	1.72784	1.50249	1.71689	1.50904	1.68884	1.50249	1.59469	1.52453	1.76378
AP_STD	0.41727	0.27487	0.45697	0.29295	0.47537	0.28245	0.45697	0.36906	0.44798	0.31383
AN_STD	0.43871	0.32888	0.44504	0.34455	0.44953	0.34881	0.44504	0.40005	0.42488	0.31295
AN_CS	0.46378	0.45772	0.40799	0.45264	0.41154	0.46871	0.40799	0.46011	0.40983	0.43755
AP_CS	0.63418	0.77936	0.56723	0.76373	0.5776	0.78183	0.56723	0.71756	0.58192	0.757
EER	0.21	**0.04794**	0.245	**0.0554**	0.239	**0.05327**	0.245	**0.12074**	0.215	**0.04545**
AUC	0.85928	0.98612	0.83398	0.9829	0.83937	0.98394	0.83398	0.94862	0.85951	0.98668

Results

Authors	Method	EER
This Paper	GAFMAT	0.04545
	GASF	0.0554
	GADF	0.04794
	RP	0.05327
	MTF	0.12074
Killourhy (original) [11]	Manhattan (scaled)	0.096
Zhong et al. [8]	Nearest Neighbor (new distance metric) + outlier removal	0.084
Zhong et al. [8]	Nearest Neighbor (new distance metric)	0.087
Monaco et al. [36]	Inductive transfer encoder (Manhattan distance)	0.063
Hayreddin et al. [35]	Convolutional Neural Network	0.065
Ivannikova et al. [53]	Dependence Clustering with Manhattan	0.077
Sae-Bae et al. [54]	Manhattan (scaled with standard deviation)	0.0916

GAFMAT Method Validation Using the GREYC Dataset



- In the initial phase of the experiments section, using the CMU dataset, we empirically established that our proposed method, called GAFMAT, provides the lowest EER value.

- The GREYC-NISLAB dataset was collected in 2013 and includes five passwords entered by 110 users. The passwords are as follows:
  - 1. "leonardo dicaprio”
  - 2. "the rolling stones"
  - 3. "michael schumacher”
  - 4. "red hot chilli peppers"
  - 5. "united states of america”

- The dataset of a single password consists of 2200 samples. In total, the dataset consists of 11000 data samples corresponding to 110 users, 20 samples per user.

GAFMAT Method Validation Using the GREYC Dataset

	GREYC_dataset				
	united states of america	leonardo di caprio	michael shumacher	the rolling stones	red hot chili peppers
	GAFMAT	**GAFMAT**	**GAFMAT**	**GAFMAT**	**GAFMAT**
Accuracy	0.99219	0.97656	0.99219	0.98698	0.97778
AP_ED	0.39566	0.44736	0.39958	0.43986	0.45165
AN_ED	1.61275	1.55644	1.48864	1.61202	1.63478
AP_STD	0.19676	0.24318	0.20467	0.21992	0.21505
AN_STD	0.38013	0.40601	0.38351	0.37381	0.38917
AN_CS	0.4979	0.49905	0.52795	0.48703	0.47839
AP_CS	0.80217	0.77632	0.80021	0.78007	0.77417
EER	**0.04688**	0.07552	0.0651	**0.04688**	**0.04444**
AUC	0.98847	0.97824	0.98771	0.98667	0.98272
	TESTING DATA				
Accuracy	0.92	0.84	0.86	0.86	0.84
AP_ED	0.75085	0.78894	0.67407	0.86642	0.8767
AN_ED	1.50073	1.55808	1.33055	1.49985	1.55131
AP_STD	0.43587	0.41371	0.31141	0.40861	0.44201
AN_STD	0.42794	0.40956	0.49554	0.41111	0.40963
AN_CS	0.43711	0.41324	0.49884	0.40843	0.393
AP_CS	0.62458	0.60553	0.66297	0.56679	0.56165
EER	0.14	0.16	0.22	0.2	0.22
AUC	0.8924	0.9032	0.854	0.8592	0.8668

Kito pusmečio darbo planas

- Publikacija mokslo leidinyje, turinčiame cituojamumo rodiklį Clarivate Analytics Web of Science duomenų bazėje (planuojama iki 2023 m. gruodžio 30 d.);
- Parengti publikaciją mokslo leidiniui, turinčiame cituojamumo rodiklį Clarivate Analytics Web of Science duomenų bazėje (planuojama iki 2024 m. balandžio mėn.);
- 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