



**Vilnius
University**

Doktoranto Karolio Šablausko ataskaita už
2024/2025 mokslo metų antrąjį pusmetį

Darbo vadovė: prof. Audronė Jakaitienė

Disertacijos pavadinimas: Genetinių pokyčių charakterizavimas naudojant giliojo mokymo neuroninius tinklus (angl. *Characterization of genetic changes using deep neural networks*)

Darbo vadovas: prof. Audronė Jakaitienė

Doktorantūros pradžios ir pabaigos metai: 2022 – 2027
(akademinės atostogos 2024-10-01 – 2025-01-31)

Studijų metai: 3.

Visų
doktorantūros
studijų ir
mokslinių tyrimų
planas bei jo
vykdymo
suvestinė

1 lentelė: Doktorantūros studijų planas

Studijų metai	Egzaminai	
	Planas	Įvykdyta
I (2022/2023)	1	1
II (2023/2024)	2	2
Akademinės atostogos 2024-10-01 – 2025-01-31		
III (2025/2026)	1	0
IV (2026/2027)	0	0
Iš viso:	4	3

Studijų metai	Dalyvavimas konferencijose				Publikacijos			
	Tarptautinėse		Nacionalinėse		Su citavimo rodikliu			Be citavimo rodiklio
	Planas	Įvykdyta	Planas	Įvykdyta	Planas	Įvykdyta	Būklė	Publikacijų be citavimo rodiklio teikti nenumatoma
I (2022/2023)	0	0	0	0	0	0	0	
II (2023/2024)	1	0	0	0	0	0	0	
Akademinės atostogos 2024-10-01 – 2025-01-31								
III (2025/2026)	1	1 pristatymas +1 skola iš II metų, pateikta santrauka, laukiama atsakymo)	0	0	1	0	1 ruošiamas rankraštis + 1 publikacija pateikta	
IV (2026/2027)	1	0	0	0	1	0		
Iš viso:	3	1	0	0	2	0		

Informacija apie tarptautinius renginius ir publikacijas (1)

Pristatytas žodinis pranešimas tarptautinėje konferencijoje „International Vilnius Conference on Statistics and Its Applications 2025“

Pranešimo tema: „Variational Autoencoder as Dimensionality Reduction Technique for Single-Cell RNA Sequencing Data“

FIRST INTERNATIONAL VILNIUS CONFERENCE ON STATISTICS AND ITS APPLICATIONS

Variational Autoencoder as Dimensionality Reduction Technique for Single-Cell RNA Sequencing Data

Karolis Sablauskas[†]

[†]Institute of Data Science and Digital Technologies, Vilnius University, Vilnius, Lithuania

Informacija apie tarptautinius renginius ir publikacijas (2)

Pateiktas abstraktas į tarptautinę konferenciją „American Society of Hematology Annual Meeting 2025“, į mašininio mokymo pranešimų dalį

Santraukos pavadinimas: „Machine learning framework for molecular classification of hematologic malignancies using transcriptome data“

BACKGROUND

Accurate characterization of hematologic malignancy is a first step in correct and tailored treatment. Tumor whole transcriptome sequencing (WTS) has recently emerged as a universal technique allowing for the accurate identification of not only all possible fusion transcripts, but also point mutations, copy number alterations, and gene overexpression, being especially useful for diagnosing cases of B-cell acute lymphoblastic leukemias (ALL). Since WTS provides gene expression landscape of the tumor sample, we decided to investigate whether the expression signatures could be used to also accurately classify a full spectrum of pediatric hematologic malignancies.

METHODS

We selected hematologic disorders with a minimum of 3 samples per condition in St. Jude Children's Research Hospital and Princess Maxima Center for Pediatric Oncology repositories. In total, this yielded WTS count data for 2914 samples across 69 distinct hematologic malignancy subtypes. The conditions spanned acute myeloid and lymphoid leukemias, Hodgkin and non-Hodgkin lymphomas, myelodysplastic syndrome, chronic myeloid leukemia, and others, including rare entities such as acute myeloid leukemia with *KMT2A*-partial tandem duplication or T-ALL with *TLX1* activation.

Raw gene counts of WTS data were used as model inputs. To focus on relevant genes in tumorigenesis, a list of 1245 known tumor drivers was selected based on OncoKB and CosmicDB. To improve classifier performance, gene counts were enriched with statistical features (total sum, mean, kurtosis and skew). Different machine learning models (linear regression, random forest, support vector classifier, multilayer perceptron and 3 types of gradient boosting models) were trained using a 3-fold cross-validation (due to minimum of 3 samples per condition). F1 score (harmonic mean of precision and recall) was used as a primary metric to account for the imbalance of different hematologic conditions. Finally, external model validation was performed on an independent dataset of 28 hematological malignancy samples of 9 conditions from Children's Clinical University Hospital, Riga, Latvia.

RESULTS

The best performing model was a voting ensemble combining 7 different classifiers with internal cross-validation accuracy of 0.82 and an F1 score of 0.59. Performance on the independent test set was similar with overall accuracy of 0.82 and F1 score of 0.53. WTS in the test set identified all fusions that were also detected by the standard method. The model additionally detected clinically relevant events such as a case of B-ALL with *PAX5-JAK2* fusion which was missed by the standard method.

Conditions having similar expression patterns such B-ALL with *ETV6-RUNX1* fusion and B-ALL *ETV6-RUNX1*-like were the most challenging to distinguish. Notably, acute megakaryoblastic leukemia cases, characterized by *HOXA9* gene rearrangements, were classified as *KMT2A*-rearranged AML likely due to the shared transcriptional activation of *HOXA9*.

Informacija apie tarptautinius renginius ir publikacijas (3)

Vilnius
University

Atnaujintas publikacijos „High-throughput single cell -omics using semi-permeable capsules“ rankraštis, kartu su autoriais iš VU Gyvybės mokslo centru.

Pateiktas publikacijai į žurnalą „Science“ 5-metų cituojamumo rodiklis 49,7. Prieš tai rankraštis teiktas žurnalui „Nature“, tačiau nebuvo priimtas

Buvo atlikti papildomi recenzentų rekomenduoti eksperimentai.

Title: High-throughput single cell -omics using semi-permeable capsules

Authors: Denis Baronas¹, Simonas Norvaisis¹, Justina Zvirblyte¹, Greta Leonaviciene¹, Vincenta Mikulenaite¹, Karolis Goda¹, Vytautas Kaseta⁴, Karolis Sablauskas^{2,3}, Laimonas Griskevicius², Simonas Juzenas¹ and Linas Mazutis^{1,5, *}

Affiliations:

¹ Institute of Biotechnology, Life Sciences Center, Vilnius University; Vilnius, Lithuania.

² Hematology, Oncology and Transfusion Medicine Center, National Cancer Center, Vilnius University Hospital Santaros Clinics; Vilnius, Lithuania.

³ Institute of Data Science and Digital Technologies, Vilnius University; Vilnius, Lithuania.

⁴ State Research Institute Centre for Innovative Medicine, Department of Stem Cell Biology; Vilnius, Lithuania.

⁵ Department of Molecular Biology, Umea University; Umea, Sweden.

Stažuotė Rygos Vaikų klinikinėje universitetinėje ligoninėje, genetikos skyriuje (2025-05):

- Išmokta apdoroti viso transkriptomo duomenis navikų diagnostikai.
- Įdiegti literatūroje aprašyti navikų klasifikacijos algoritmai naudojant viso transkriptomo duomenis.
- Įvertinti literatūroje aprašytų navikų klasifikacijos algoritmų trūkumai.
- Surinkti ir paruošti nauji duomenys išorinei algoritmo validacijai.
- Pirminiai eksperimentai apmokant naują klasifikacijos algoritmą – vėliau pratęsta grįžus iš stažuotės.
- Iš surinktų duomenų pateiktas abstraktas konferencijai bei rengiamas publikacijos rankraštis.

Visų doktorantūros mokslinių tyrimų ir disertacijos rengimo etapai

2.3. <u>Empirinis tyrimas:</u> 1. Siūlomų algoritmų pritaikymas naujos kartos sekoskaitos duomenims. 2. Siūlomų algoritmų tobulinimas, atsižvelgiant į gautus rezultatus.	2025 m. vasario mėn. – 2025 m. birželio mėn.	Surinkti duomenys modelio apmokymui (St Jude ir Princess Maxima ligoninių kohortos) ir išorinei validacijai (Rygos vaikų universitetinės ligoninės duomenys). Išbandyti skirtingi mašininio mokymo metodai (atsitiktinio miško, logistinės regresijos, neuroninio tinklo, gradiento didinimo algoritmai). Detaliau mokslinėje ataskaitoje.
2.4. <u>Gautų duomenų analizė, apibendrinimas, išvadų parengimas:</u> 1. Algoritmų tikslumo įvertinimas, palyginimas su kitų autorių metodais, atrinktais remiantis išanalizuota moksline literatūra. 2. Gautų rezultatų apibendrinimas. 3. Išvadų parengimas.	2025 m. birželio mėn. – 2025 m. gruodžio mėn.	Modelis patikrinimas ant išorinių duomenų, identifikuotos silpnosios modelio dalys. Modelis atnaujintas padidinus mokymo duomenų kiekį (pridėti GTEX sveikų asmenų kohortos duomenys) ir apmokytas ansamblio metodas (artimiausių kaimynų ir klasifikacijos algoritmų apjungimas). Iš duomenų hematologiniams navikams pateikta santrauka konferencijai. Iš visų navikų duomenų ruošiama publikacija. Detaliau mokslinėje ataskaitoje.

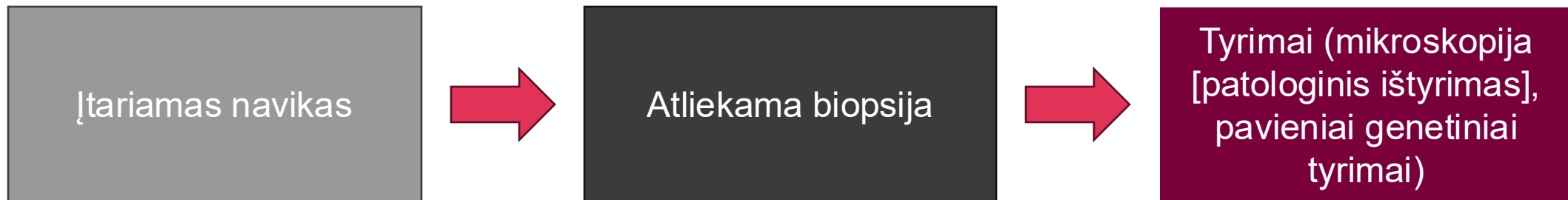
1.1 Thesis aim (Tyrimo tikslas)

To contribute to the advancement of deep learning techniques for the analysis of next generation sequencing data, with a focus on RNA sequencing data.

1.2 Thesis objectives (Tyrimo uždaviniai)

- **Data preprocessing and feature engineering:** create efficient data preprocessing pipeline suited for RNA sequencing data, including normalization and batch effect correction to prepare the data for further analysis.
- **Deep differential expression analysis:** develop and implement deep learning-based approach for identifying differentially expressed genes and pathways.
- **Evaluation and benchmarking:** conduct extensive benchmarking and cross-validation experiments to assess the performance and generalizability of deep learning models, comparing them to traditional methods.
- **Biological case study:** apply said techniques in the interpretation of biological data gathered during a biomedical study.

Mokslinių rezultatų pristatymas: tyrimo problematika



Tikslas: konkrečiai
įvardinti naviko tipą

Trukmė: 1-4 sav.

Tyrimo problematika

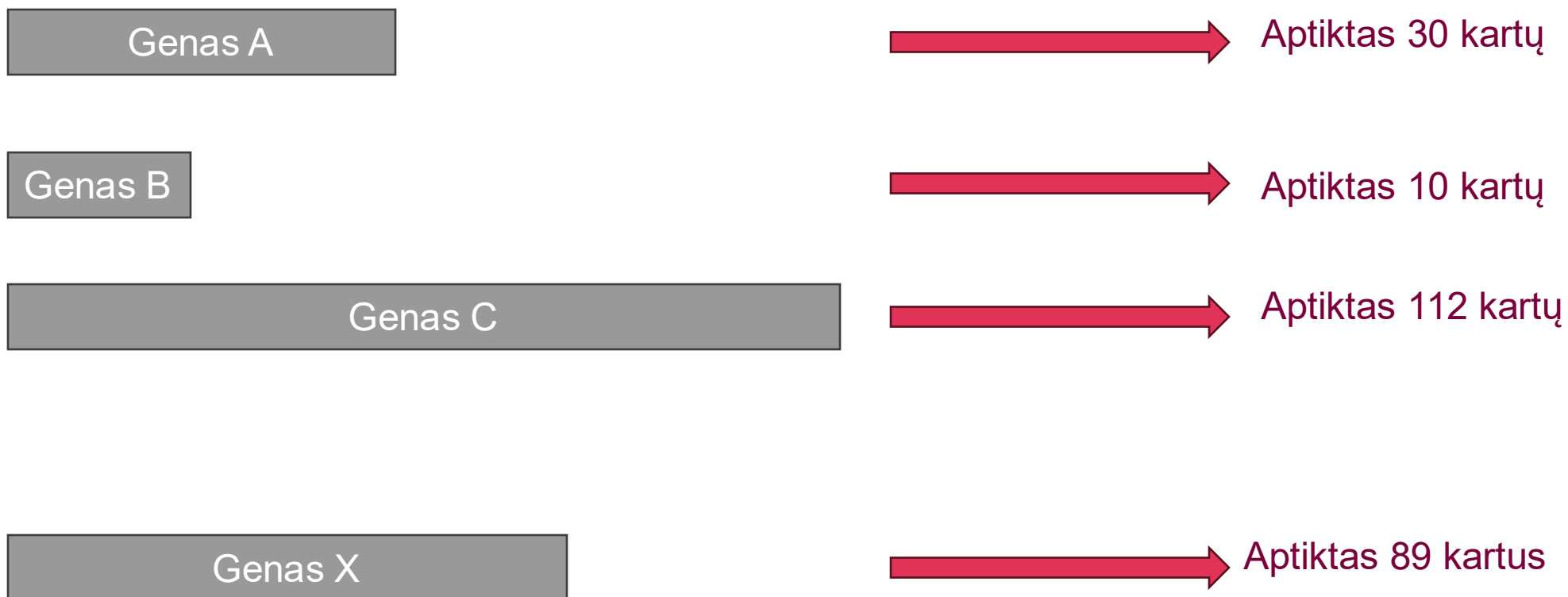
Pavienių diagnostinių tyrimų trūkumai:

- Sudėtinga iš infrastruktūros pusės (kiekvienam navikui reikia atskiro lab. metodo)
- Laiko trukmė gauti rezultatą.

Galimas sprendimas:

- “omikos” metodai – didelio kiekio genų ištyrimas

Viso transkriptomo sekoskaita



3.8.1 Task definition

The task is to train a multiclass classification model with a supervised approach. The input is a vector of size G , where G corresponds to the number of input genes:

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_G \end{bmatrix} \in \mathbb{R}^G$$

where x_i denotes the normalized expression value of the i -th gene. The corresponding target variable y is a categorical label indicating the tumor type, such that $y \in \{1, 2, \dots, C\}$ for C possible classes ($C=160$ tumor types).

Duomenys

Mokymo ir kryžminės validacijos

St. Jude ligoninė
STJUDE
(n=4395)

Princess Maxima
ligoninė
PM (n=1042)

Sveikų asmenų
kohorta
GTEX (n=803)

Išorinio testavimo

Rygos vaikų ligoninė
BKUS (n=112)

Rygos universitetas
RSU (n=123)

Testavimo

Be hipotezės (*exploratory*)

Pacientai be aiškio diagnozės
(STJUDE, PM, BKUS n=784)

Vidinio testavimo

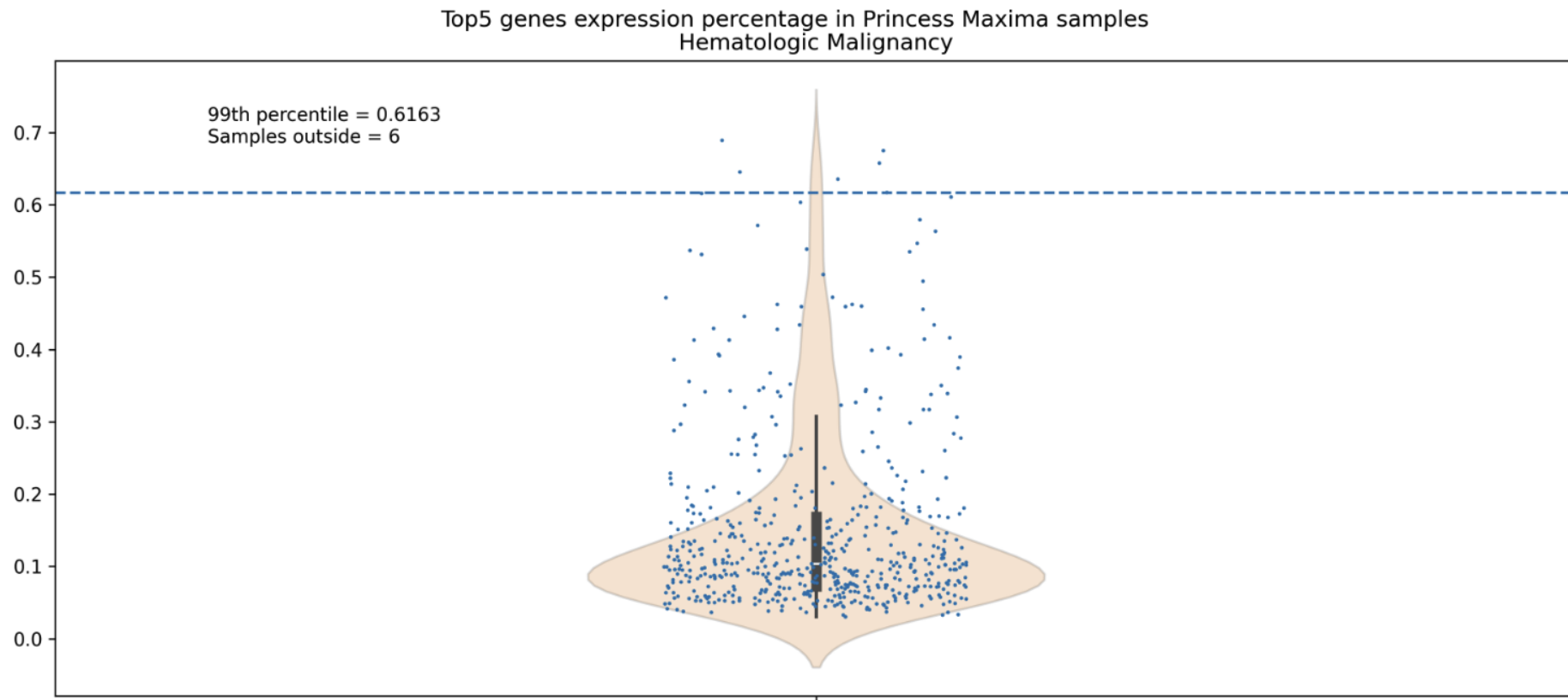
Princess Maxima
ligoninė testavimo
kohorta
PM test (n=435)

**Vilnius
University**

Kokybės kontrolė

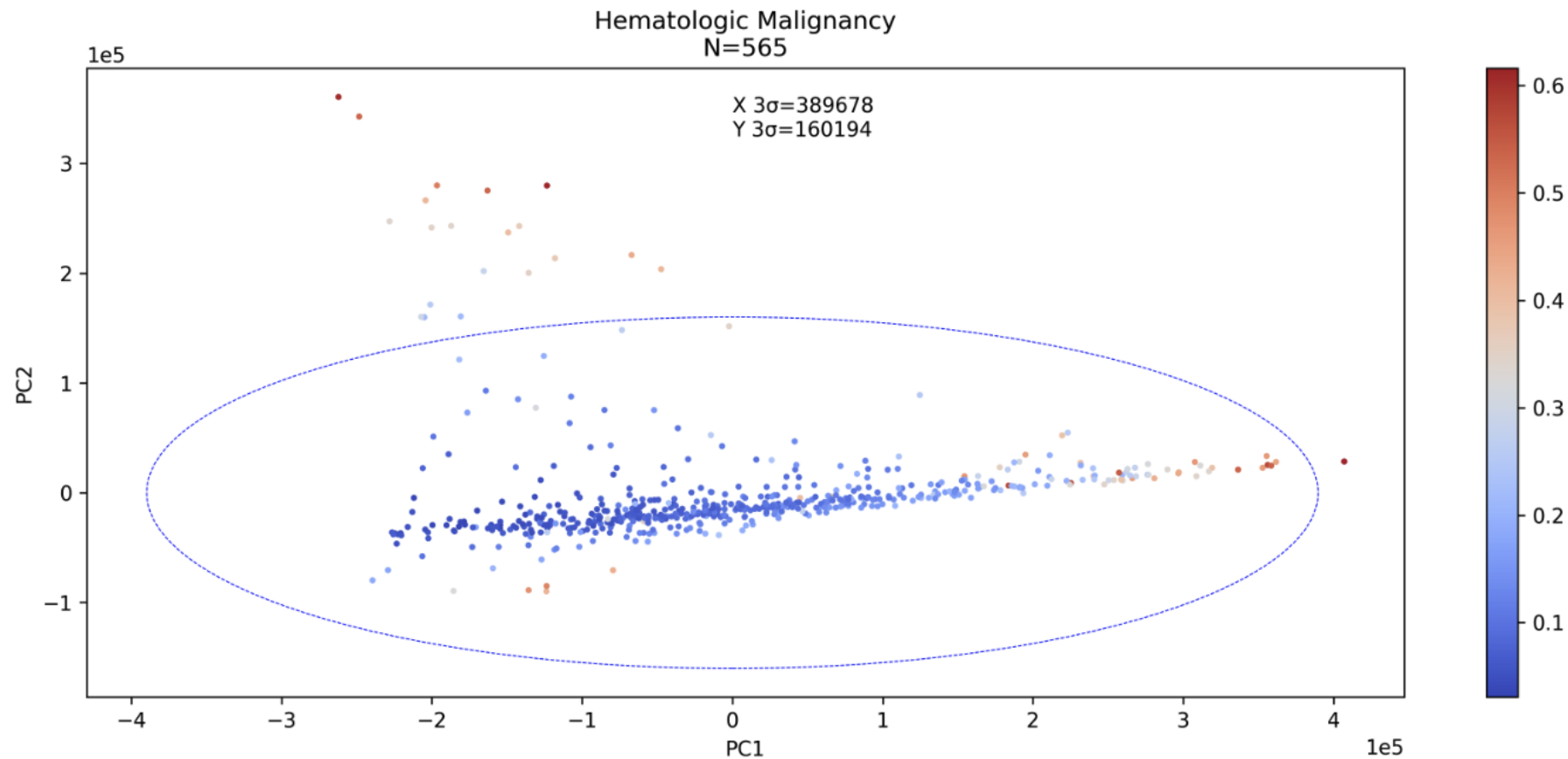
Kokybės kontrolės įvertis:
kokia dalis visų transkriptų
(geno perskaitymų) gauta
iš top 5 didžiausio
perskaitymo genų.

99-a procentilė naudojama
kaip riba.



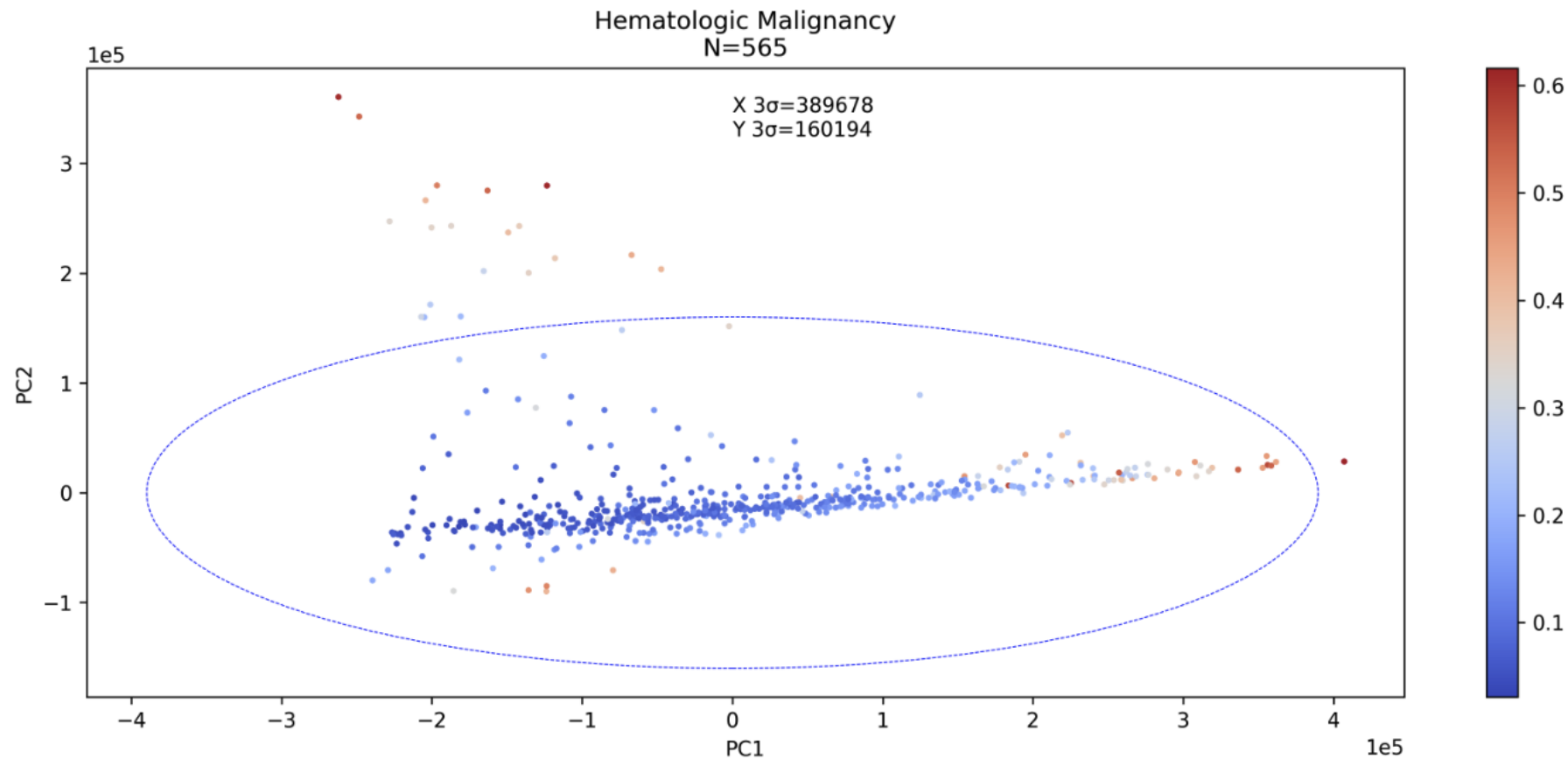
Kokybės kontrolė

Pagrindinių komponentų analizė – per daug nutolę mėginiai nuo centro - prastesnė kokybė.



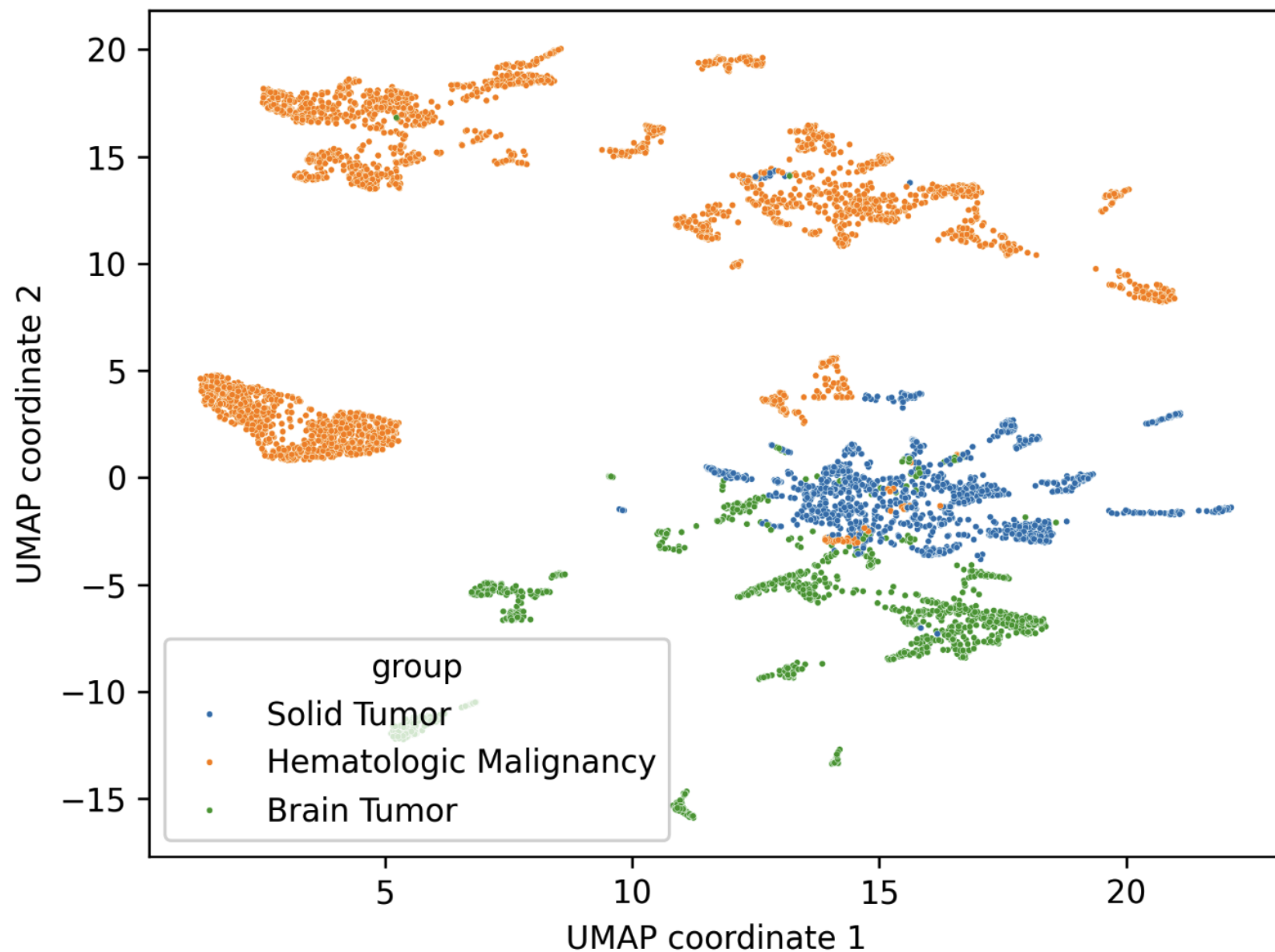
Kokybės kontrolė

Pagrindinių komponentų analizė – per daug nutolę mėginiai nuo centro - prastesnė kokybė.



Ar įmanoma klasifikuoti?

Atlikus dimensijų sumažinimą galima matyti, jog susidaro atskiri navikų klasteriai, tačiau yra tikrų persiklojimų.



Rezultatai - kryžminės validacijos duomenys

3.8.6 Performance on cross-validation data

The table below shows performance for the pan-cancer classifier accross all categories.

Classifier	F1	Accuracy	Precision	Recall
MLP	0.58	0.84	0.61	0.59
RF	0.44	0.82	0.47	0.44
LR	0.56	0.84	0.58	0.56
SVC	0.34	0.75	0.40	0.33
XGB	0.49	0.81	0.52	0.50
CatBoost	0.41	0.78	0.45	0.40
TabNet	0.52	0.79	0.54	0.54

Table 5: Performance of different models for the classification of tumor type across all 160 tumor types on cross-validation (training set).

Rezultatai— išorinės validacijos duomenys

RSU – sveiki asmenys,
sudėtingesni algoritmai
šiems asmenims spėja
naviko buvimą.

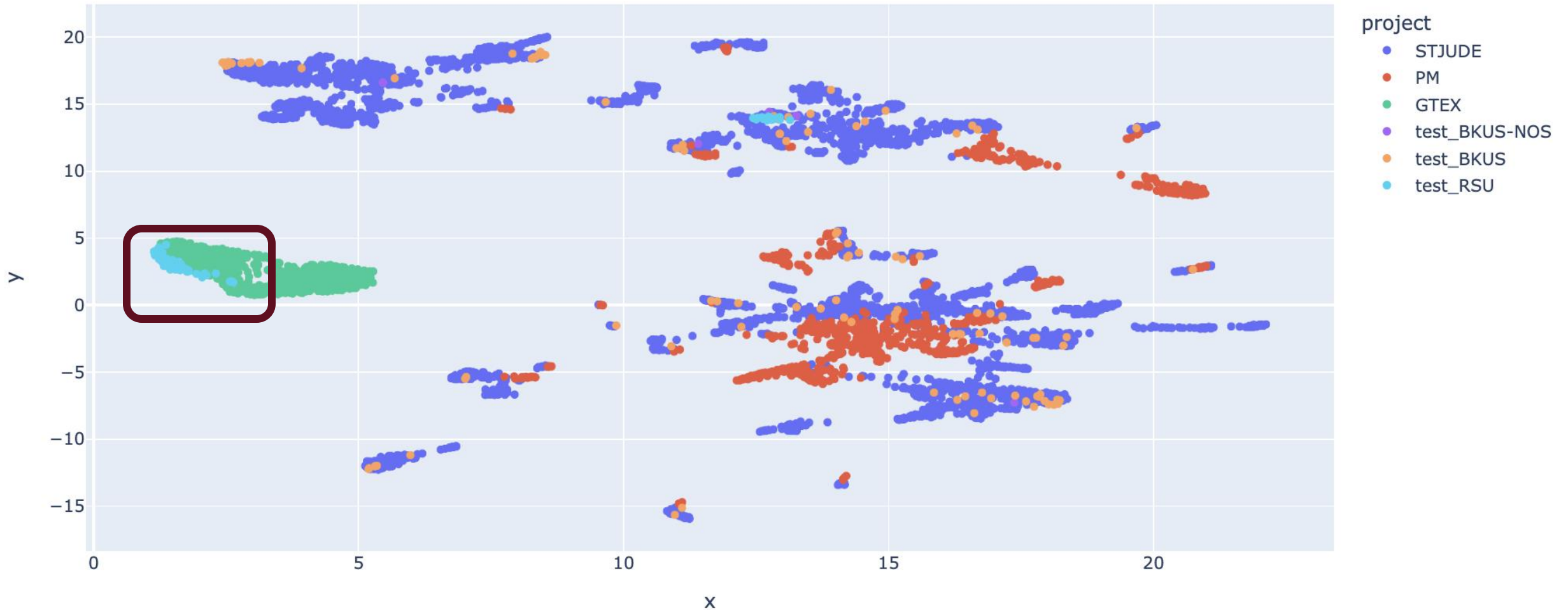
3.8.7 Performance on test data

The table below shows performance for the pan-cancer classifier accross all categories on the three testing datasets with known ground-truth labels.

Dataset	Classifier	F1	Accuracy	Precision	Recall
BKUS	MLP	0.51	0.76	0.53	0.52
PM	MLP	0.59	0.80	0.64	0.59
RSU	MLP	0.04	0.12	0.17	0.02
BKUS	RF	0.45	0.77	0.44	0.49
PM	RF	0.51	0.77	0.57	0.49
RSU	RF	0.50	0.99	0.50	0.50
BKUS	LR	0.55	0.80	0.56	0.57
PM	LR	0.54	0.77	0.59	0.55
RSU	LR	0.07	0.30	0.14	0.04
BKUS	SVC	0.42	0.71	0.42	0.45
PM	SVC	0.21	0.34	0.33	0.18
RSU	SVC	0.15	0.30	0.33	0.10
BKUS	XGB	0.49	0.73	0.49	0.54
PM	XGB	0.45	0.73	0.49	0.46
RSU	XGB	0.01	0.10	0.07	0.01
BKUS	CatBoost	0.53	0.79	0.53	0.54
PM	CatBoost	0.36	0.64	0.38	0.40
RSU	CatBoost	0.02	0.07	0.12	0.01
BKUS	TabNet	0.49	0.73	0.49	0.53
PM	TabNet	0.32	0.58	0.37	0.33
RSU	TabNet	0.02	0.08	0.14	0.01

Table 6: Performance of different models for the classification of tumor type across all 160 tumor types on three test sets.

— Išorinės validacijos duomenų projekcija į UMAP



Rezultatai— išorinės validacijos duomenys

RSU – sveiki asmenys,
sudėtingesni algoritmai
šiems asmenims spėja
naviko buvimą.

Dataset	Classifier	F1	Accuracy	Precision	Recall
BKUS	MLP	0.52	0.79	0.54	0.53
PM	MLP	0.58	0.78	0.64	0.57
RSU	MLP	0.23	0.85	0.25	0.21
BKUS	RF	0.55	0.80	0.56	0.56
PM	RF	0.36	0.66	0.45	0.35
RSU	RF	0.23	0.85	0.25	0.21
BKUS	LR	0.59	0.82	0.61	0.61
PM	LR	0.54	0.75	0.59	0.54
RSU	LR	0.23	0.85	0.25	0.21
BKUS	SVC	0.57	0.80	0.58	0.58
PM	SVC	0.35	0.64	0.43	0.34
RSU	SVC	0.23	0.85	0.25	0.21
BKUS	XGB	0.59	0.82	0.59	0.62
PM	XGB	0.42	0.74	0.49	0.42
RSU	XGB	0.18	0.85	0.20	0.17
BKUS	CatBoost	0.55	0.80	0.55	0.57
PM	CatBoost	0.39	0.67	0.45	0.39
RSU	CatBoost	0.18	0.85	0.20	0.17
BKUS	TabNet	0.56	0.81	0.55	0.60
PM	TabNet	0.38	0.67	0.44	0.38
RSU	TabNet	0.15	0.86	0.17	0.14

Table 7: Performance of different models for the classification of tumor type across all 160 tumor types with addition of KNN to each classifier on three test sets

Naujumas

- Kokybės kontrolės žingsniai duomenų atrankai ir požymių atranka
- Aptinkamų navikų skaičius (160 vs 96 prieš tai aprašyti literatūroje)
- Klasifikatorių kombinacija
- Testavimo duomenys (kitos publikacijos neturėjo išorinės validacijos)

Kito semestro planas

- Egzaminas 4/4
- Publikacijos pateikimas 1/2
- Pranešimo skaitymas tarptautinėje konferencijoje