Analiza genske ekspresije strojnim učenjem u svrhu identifikacije tumora

Dokumentacija
Verzija 1.0

Student: Luka Bulić

Nastavnik: izv. prof. dr. sc. Mirjana Domazet-Lošo

Sadržaj

1. Uvod

Neoplazme nepoznatog primarnog sjela predstavljaju značajan dijagnostički izazov u kliničkoj onkologiji. Kada se susretnu s ovim izazovom, liječnici se načešće okrenu dijagnostičkim metodama poput PET-CT pretrage, u nadi da će njima detektirati primarno sjelo. Međutim, ove su se metode u literaturi pokazale neadekvatnima za značajan broj pacijenata, kod kojih primarno sjelo ostane nedijagnosticirano [1].

Alternativni pristup koji se trenutno istražuje obuhvaća metode molekularne dijagostike koje bi potencijalno mogle suziti mjesto traženja. Specifično, genetskom profiliranju tumora se u zadnje vrijeme pridaje velika pažnja u znanstvenoj zajednici. Određene studije koje utiliziraju genetsko tumorsko profiliranje nastoje razriješiti problem nepoznatog primarnog tumorskog sjela kroz istraživanje korelacija između genetskih obilježja i određenih vrsta tumora, koristeći tradicionalne metode statističke analize ili novije pristupe temeljene na algoritmima strojnog učenja [2, 3].

U ovom istraživanju, cilj je bio identificirati tipove i podtipove tumora koristeći obilježja genetske ekspresije i modele strojnog učenja. Ovim smo pristupom evaluirali korisnost dijagnostike DNA mikročipovima u identificiranju originalnog tumorskog sjela. Nadalje, određena su najsignifikantnija genetska obilježja za svaki tip tumora te su ti rezultati interpretirani u kontekstu javno dostupne baze podataka kako bi se razjasnili mehanizmi razlikovanja pojedinih tipova.

2. Metode

2.1. Opis podataka

Analizirali smo uređenu CuMiDa bazu podataka, izrađenu na temelju podataka o ekspresiji Affymetrix proba u zdravom i tumorskom tkivu [4]. Podskupovi podataka korišteni u istraživanju bili su “Brain_GSE50161”, “Leukemia_GSE28497”, “Liver_GSE14520_U133A”, “Lung_GSE19804” i “Renal_GSE53757”. Podskupovi su spojeni te su zadržane isključivo zajedničke značajke svih 5 podskupova. Definirane klase su: „Brain“, „Leukemia“, „Liver“, „Lung“, „Renal“ i „Normal“, koje su se slijedno odnosile na neoplazme središnjeg živčanog sustava, jetre, pluća, bubrega, kvrnih stanica te zdrava tkiva. Klasa „Brain“ imala je 4 podtipa, a klasa „Leukemia“ 7 podtipova, dok su ostale klase imale po 1 podtip.

2.2. Treniranje modela i odabir značajki

Analize i vizualizacija podataka napravljeni su koristeći postojeće Python knjižnice (Pandas, SciKit-Learn i XGBoost) [5-8]. Reproducibilnost rezultata osigurana je kontrolom nasumičnih varijabli. Tipovi i podtipovi tumora klasificirani su korišteni algoritam slučajne šume (RFC) te XGBoost (XGB) algoritam. Nadalje, zasebni RFC modeli trenirani su za klasifikaciju

podtipova klasa „Leukemia“ i „Brain“. Deset najsignifikantnijih značajki određeno je za svaku klasu, osim „Normal“, koristeći kompleksnu funkciju koja kombinira korelacijsku metodu (CM), info gain metodu (IGM) te ExtraTrees algoritam za klasifikaciju (ETC). Baza podataka je podešena za svaki tip tumora, svrstavajući sve ostale tipove u klasu „Ostalo“, kako bi se problem reducirao na binarnu klasifikaciju. Svaka od 3 metode producirala je 50 najznačajnijih značajki za svaku klasu, nakon čega su probrane one koje su odredile barem dvije od tri metode. Konačno, iz tog je skupa preuzeto deset značajki.

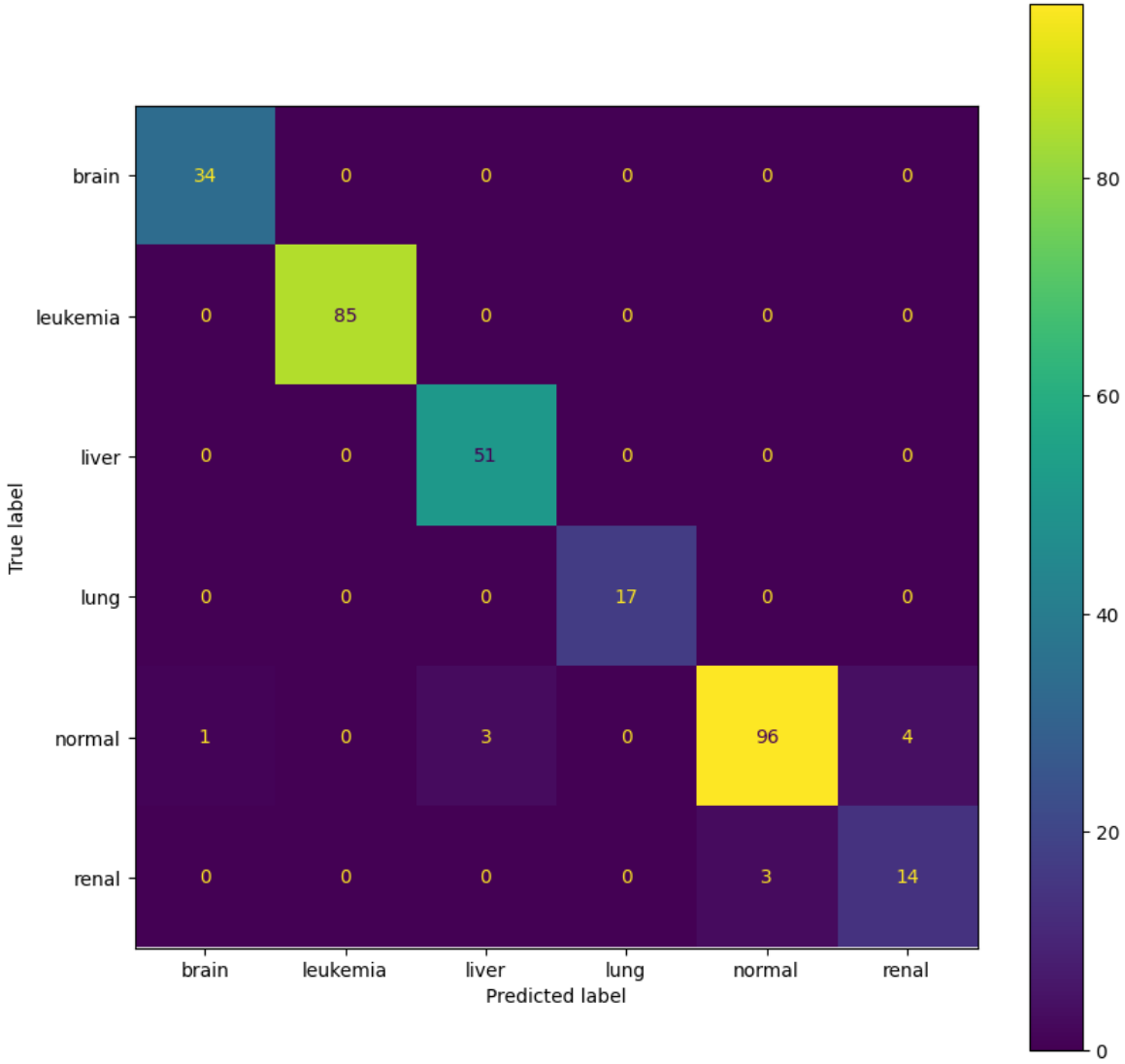
2.3. Evaluacija modela

Modeli za klasifikaciju tipa i podtipa tumora evaluirani su na temelju točnosti, matrica zabune te *receiver operating characteristic* (ROC) analize. Glede selekcije značajki, prosječne vrijednosti ekspresije (na logaritamskoj skali) bile su izračunate za svaku odabranu značajku i uspoređene s referentnim prosječnim vrijednostima iz *Gene Enrichment Profiler* (GEP) baze podataka [9]. Prag za značajnu ekspresiju definiran je u GEP bazi te iznosi 6,65 ($\log_2 100$). Podudarnost praga definirana je za značajke koje su prešle prag u našoj i referentnoj bazi ili niti u jednoj. Specifično, u slučaju nepodudarnosti zbog preniske ekspresije u referentnoj bazi, razmatrana je vrijednost značajke iz referentne baze u zdravom tkivu. Razlike između naše i referentne baze određene su analizom podudarnosti praga Hi-kvadrat statističkim testom s Yates korekcijom.

3. Rezultati

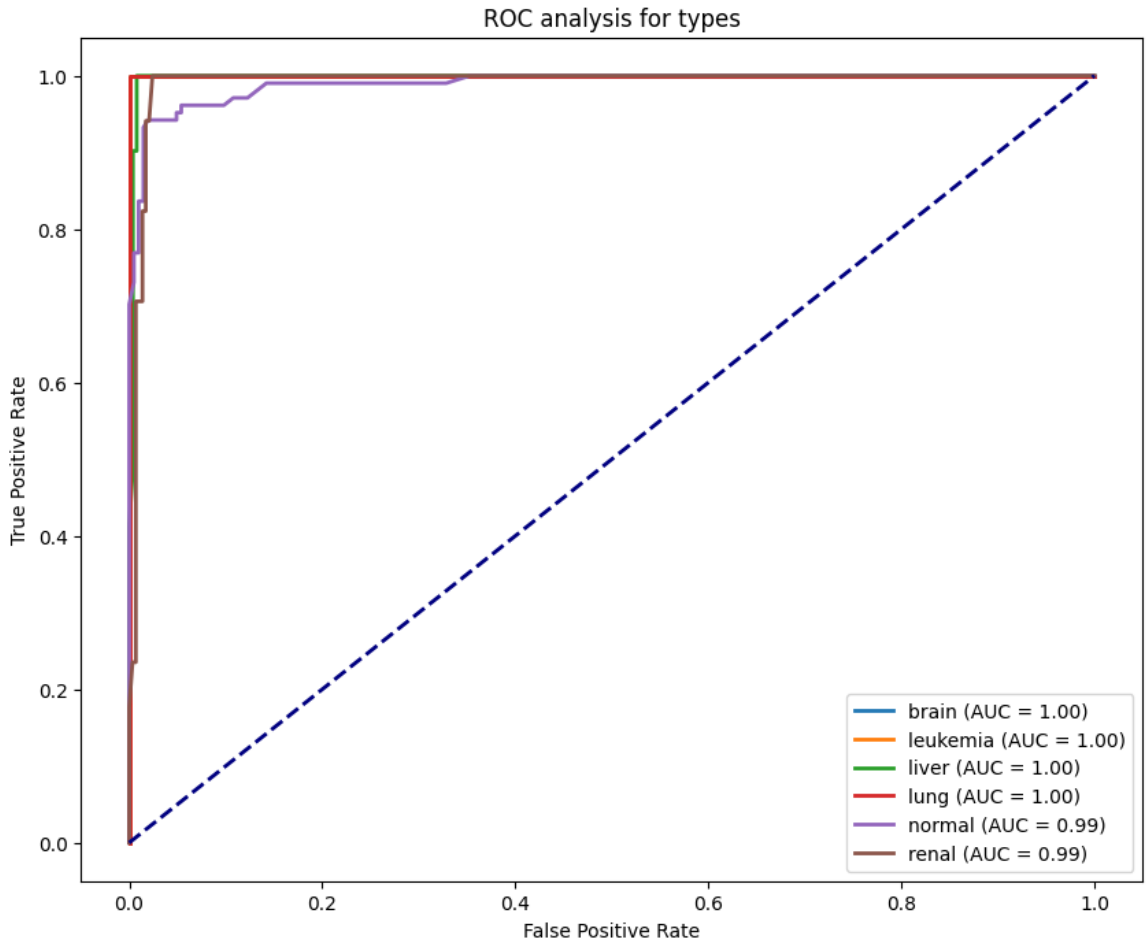
3.1. Analiza kvalitete

Nakon pretprocesiranja, baza podataka sadržavala je 1025 uzoraka i 22277 značajki genetske ekspresije. RFC klasifikacijski modeli pokazali su točnosti od 96,4% za tipove tumora i 90,3% za podtipove tumora. Matrica zabune RFC modela za klasifikaciju tumora pokazala je visoke točnosti (≥90%) za svaki tip osim za „Renal“ klasu (82,4%).



Slika 1. RFC model za klasifikaciju tipova – matrica zabune.

ROC analiza RFC modela za klasifikaciju tipova pokazala je vrijednost površine ROC krivulje ≥0,99 za svaki tip.



Slika 2. RFC model za klasifikaciju tipova – ROC analiza.

Zasebni RFC modeli trenirani za identifikaciju podtipova leukemija i tumora središnjeg živčanog sustava imali su slijedno točnosti 85,9% i 89,7%. S druge strane, XGB modeli imali su niže točnosti, 95,1% za klasifikaciju tipova i 89,3% za klasifikaciju podtipova.

Potencijal za identifikaciju tumora na temelju DNA mikročipova demonstriran je u studiji objavljenoj prije 20 godina [10]. Ovaj projekt ne samo potvrđuje taj potencijal nego nudi i praktičnu metodu utilizacije podataka u tu svrhu. Prije usporedbe s modelima koji su već objavljeni u literaturi, ograničenja ovakvih usporedbi moraju se razmotriti. Ona postoje zbog razlike u broju klasa, razlike u uzorcima i značajkama te razlike u kvaliteti podataka, razlika koje su visoko varijabilne među različitim studijama. Primjena umjetne inteligencije u analizi RNA ekspresije za identifikaciju tumora rađena je koristeći kompleksne modele poput konvolucijskih neuralnih mreža. U smislu točnosti, naš model dostiže razinu kvalitete onih već objavljenih u literaturi [11]. Drugi aspekti identifikacije tumorskog podrijetla bazirani su na modelima umjetne inteligencije poput modela dubokog učenja koji analiziraju histološke presjeke te modela za analizu somatskih mutacija. U usporedbi s ovim modelima, naš model također drži dobru točnost [12, 13]. Strojno učenje također je primjenjeno u analizi PET-CT nalaza za druge predikcije vezane uz tumore. Međutim, u smislu tumora nepoznatog primarnog sjela, daljnja istraživanja u ovom smislu su potrebna [14].

Pri razmatranju ekonomskog aspekta, DNA mirkočipovi imaju veliku prednost nad naprednim tehnologija genomskog sekvenciranja, s obzirom da su jeftiniji, brži i dostupniji [15]. S druge strane, u usporedbi s PET-CT-om, ekonomski aspekt težak je za procijeniti. Detaljna analiza koja bi uzela u obzir točnosti PET-CT-a i naših modela u razrješenju tumora nepoznatog primarnog sjela te rezultirajućih razlika u troškovima liječenja morala bi biti napravljena u svrhu ove procjene.

3.2. Analiza odabira značajki

Kombinirana funkcija za odabir značajki uspješno je odredila deset najsignifikantnijih značajki za svaki tip tumora. Usporedba s GEP bazom podataka nije pokazala statistički signifikantnu razliku. Analiza podudarnosti praga Hi-kvadrat testom s Yates korekcijom pokazala je p-vrijednost >0,1 za svaku klasu. Primjetna je razlika uočena među klasama glede izvornog tkiva na temelju kojeg je uočena podudarnost praga. Podudarnost za klase „Liver“ i „Leukemia“ bazirana je uglavnom na vrijednostima iz tumorskog tkiva. S druge strane, podudarnost za klase „Brain“, „Lung“ i „Renal“ bazirana je uglavnom na vrijednostima iz zdravog tkiva. Ova diskrepancija zahtijeva daljnje istraživanje.

Tablica 1. Usporedba ekspresije najsignifikantnijih značajki s vrijednostima iz GEP baze podataka

PROBA	GEN	Prosječna ekspresija	Referentna ekspresija (TUMOR)	Referentna ekspresija (ORGAN)*	Podudarnost praga (T≈6,65)
JETRA					
204046_at	*PLCB2*	=4.10 (-)	≈2.50 (-)	/	Y
207667_s_at	*MAP2K3*	=4.85 (-)	≈7.25 (+)	/	N
209365_s_at	*ECM1*	=5.21 (-)	≈3.25 (-)	/	Y
215178_x_at	*MPV17L*	=4.12 (-)	≈7.00 (+)	/	N
215712_s_at	*IGFALS*	=4.10 (-)	≈2.75 (-)	/	Y
204018_x_at	*HBA2*	=7.14 (+)	≈3.00 (-)	≈12.50 (+)	Y_O
209889_at	*SEC31B*	=4.35 (-)	≈2.00 (-)	/	Y
211166_at	*FAM153C*	=3.96 (-)	≈4.00 (-)	/	Y
211745_x_at	*HBA2*	=7.41 (+)	≈4.00 (-)	≈12.00 (+)	Y_O
211768_at	N/A	=3.92 (-)	≈4.00 (-)	/	Y
p-vrijednost (Hi-kvadrat test s Yates korekcijom) = 0,626					
LEUKEMIJA					
217757_at	*A2M*	=3.72 (-)	≈2.00 (-)	/	Y
221767_x_at	*HDLBP*	=5.53 (-)	≈8.00 (+)	/	N
221802_s_at	*KIAA1598*	=2.60 (-)	≈4.25 (-)	/	Y
201876_at	*PON2*	=3.66 (-)	≈7.50 (+)	/	N
202054_s_at	*ALDH3A2*	=4.34 (-)	≈5.50 (-)	/	Y
202090_s_at	*UQCR*	=7.06 (+)	≈10.50 (+)	/	Y
202842_s_at	*DNAJB9*	=2.76 (-)	≈6.00 (-)	/	Y
203382_s_at	*APOE*	=4.08 (-)	≈4.25 (-)	/	Y
207335_x_at	*ATP5I*	=5.40 (-)	≈10.50 (+)	/	N

208791_at	*CLU*	=3.71 (-)	≈7.50 (+)	/	N
p-vrijednost (Hi-kvadrat test s Yates korekcijom) = 0,302					
PLUĆA					
203757_s_at	*CEACAM6*	=12.75 (+)	≈2.25 (-)	≈10.75 (+)	Y_O
205941_s_at	*COL10A1*	=10.17 (+)	≈2.50 (-)	≈2.50 (-)	N
209173_at	*AGR2*	=12.15 (+)	≈5.50 (-)	≈9.00 (+)	Y_O
210608_s_at	*FUT2*	=8.31 (+)	≈2.75 (-)	≈2.75 (-)	N
211657_at	*CEACAM6*	=12.79 (+)	≈5.75 (-)	≈12.50 (+)	Y_O
217428_s_at	*COL10A1*	=8.99 (+)	≈4.00 (-)	≈3.75 (-)	N**
218186_at	*RAB25*	=10.86 (+)	≈3.00 (-)	≈7.25 (+)	Y_O
218960_at	*TMPRSS4*	=8.80 (+)	≈3.25 (-)	≈3.25 (-)	N**
219388_at	*GRHL2*	=8.31 (+)	≈2.50 (-)	≈3.25 (-)	N
37004_at	*SFTPB*	=12.61 (+)	≈2.75 (-)	≈13.75 (+)	Y_O
p-vrijednost (Hi-kvadrat test s Yates korekcijom) = 0,112					
BUBREG					
205532_s_at	*CDH6*	=8.76 (+)	≈3.50 (-)	≈4.50 (-)	N
205799_s_at	*PREPL*	=12.35 (+)	≈2.75 (-)	≈12.50 (+)	Y_O
206228_at	*PAX2*	=8.85 (+)	≈3.00 (-)	≈6.75 (+)	Y_O
207470_at	N/A	=5.52 (-)	≈2.50 (-)	/	Y
210735_s_at	*CA12*	=10.03 (+)	≈4.25 (-)	≈8.50 (+)	Y_O
215392_at	N/A	=6.75 (+)	≈3.50 (-)	≈3.75 (-)	N
219271_at	*GALNT14*	=10.56 (+)	≈4.25 (-)	≈7.00 (+)	Y_O
219621_at	N/A	=3.09 (-)	≈4.00 (-)	/	Y
221009_s_at	*ANGPTL4*	=10.68 (+)	≈3.00 (-)	≈3.50 (-)	N
222281_s_at	N/A	=9.31 (+)	≈3.50 (-)	≈6.75 (+)	Y_O
p-vrijednost (Hi-kvadrat test s Yates korekcijom) = 0,348					
SREDIŠNJI ŽIVČANI SUSTAV* **					
204966_at	*BAI2*	=9.53 (+)	N/A	≈8.50 (+)	Y_O
203540_at	*GFAP*	=12.95 (+)	N/A	≈12.00 (+)	Y_O
205493_s_at	*DPYSL4*	=9.73 (+)	N/A	≈6.75 (+)	Y_O
211842_s_at	*SLC24A1*	=6.20 (-)	N/A	≈2.00 (-)	Y_O
212843_at	*NCAM1*	=11.50 (+)	N/A	≈10.75 (+)	Y_O
214393_at	*RND2*	=7.74 (+)	N/A	≈3.50 (-)	N
214772_at	*C11orf41*	=8.70 (+)	N/A	≈5.00 (-)	N
216225_at	N/A	=5.87 (-)	N/A	≈3.00 (-)	Y_O
217366_at	N/A	=6.78 (+)	N/A	≈2.75 (-)	N
220278_at	*JMJD2D*	=6.89 (+)	N/A	≈2.50 (-)	N
p-vrijednost (Hi-kvadrat test s Yates korekcijom) = 0,171					

(+) – Ekspresija iznad praga; (-) – Ekspresija ispod praga; Y – Podudarnost praga bazirana na tumorskom tkivu; Y_O – Podudarnost praga bazirana na zdravom tkivu; N – Nepodudarnost praga; / - Nije određeno; N/A – Nedostupno

*Određeno samo ako je prosječna ekspresija bila iznad praga, a referentna ekspresija za tumorsko tkivo ispod praga

**Podudarnost praga za tkivo dušnika

***Referentne ekspresije za tumorsko tkivo nisu se mogle odrediti zbog specifičnosti

4. Ograničenja

Ovo istraživanje ima potencijalna ograničenja. Prvo, točnost modela varirala je od 2% pri različitim vrijednostima nasumičnih varijabli. Prvo odabrano nasumično stanje bilo je korišteno (n = 42) te nasumična stanja nisu korištena radi postizanja bolje točnosti. Drugo, neke vrijednosti ekspresije u GEP pokazale su visoku varijabilnost. Treće, s obzirom na jedinstvenost procesirane baze podataka, usporedbe kvalitete s ostalim studijama su otežane. Konačno, faktor kvalitete podataka iz kojih je CuMiDa baza izrađena mora biti uzet u obzir.

5. Zaključak

Rezultati istraživanja pokazuju korist podataka DNA mikročipova, kao i ovih modela, u rješavanju tumora nepoznatog primarnog sjela. Primijenjen je integrirani pristup koji kombiniran podatke DNA mikročipova i strojnog učenja. Daljnja istraživanja u ovom aspektu mogla bi dovesti do izrade algoritama vrijednog implementacije u kliničku praksu.

6. Literatura

1. Bicakci N. Diagnostic and prognostic value of F-18 FDG PET/CT in patients with carcinoma of unknown primary. North Clin Istanb. 2022;9(4):337-346.

2. Pan Z, Lin H, Fu Y, et al. Identification of gene signatures associated with ulcerative colitis and the association with immune infiltrates in colon cancer. Front Immunol. 2023;14:1086898.

3. Yin X, Wu Q, Hao Z, Chen L. Identification of novel prognostic targets in glioblastoma using bioinformatics analysis. Biomed Eng Online. 2022;21(1):26.

4. Feltes BC, Chandelier EB, Grisci BI, Dorn M. CuMiDa: An Extensively Curated Microarray Database for Benchmarking and Testing of Machine Learning Approaches in Cancer Research. J Comp Biol. 2019;26(4):376-386.

5. Van Rossum G, Drake FL. Python 3 Reference Manual. Scotts Valley, CA: CreateSpace; 2009.

6. The pandas development team. pandas-dev/pandas: Pandas. Zenodo; 2024.

7. Pedregosa et al. Scikit-learn: Machine Learning in Python. JMLR 12, pp. 2825-2830, 2011.

8. Chen T, Guestrin C. XGBoost: A Scalable Tree Boosting System. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining [Internet]. New York, NY, USA: ACM; 2016. p. 785–94.

9. Benita et. al, Gene enrichment profiles reveal T cell development, differentiation and lineage specific transcription factors including ZBTB25 as a novel NF-AT repressor. Blood. 2010. Apr 21.

10. Golub TR, Slonim DK, Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531-537.

11. Zhao Y, Pan Z, Namburi S, et al. CUP-AI-Dx: A tool for inferring cancer tissue of origin and molecular subtype using RNA gene-expression data and artificial intelligence. EBioMedicine. 2020;61:103030.

12. Lu MY, Chen TY, Williamson DFK, et al. AI-based pathology predicts origins for cancers of unknown primary. Nature. 2021;594(7861):106-110.

13. Liu X, Li L, Peng L, et al. Predicting Cancer Tissue-of-Origin by a Machine Learning Method Using DNA Somatic Mutation Data. Front Genet. 2020;11:674.

14. Sadaghiani MS, Rowe SP, Sheikhbahaei S. Applications of artificial intelligence in oncologic 18F-FDG PET/CT imaging: a systematic review. Ann Transl Med. 2021;9(9):823.

15. Suratannon N, van Wijck RTA, Broer L, et al. Rapid Low-Cost Microarray-Based Genotyping for Genetic Screening in Primary Immunodeficiency. Front Immunol. 2020;11:614.