Full title:

Predicting cancer origins with a DNA methylation-based deep neural network model

Short title:

DNN model for cancer origin prediction

Authors:

Chunlei Zheng[1] and Rong Xu[1]*

[1]Department of Population and Quantitative Health Sciences, School of Medicine, Case Western Reserve University, Cleveland, OH, USA

*Corresponding to: Rong Xu

Email: rxx@case.edu (RX)

Abstract

Cancer origin determination combined with site-specific treatment of metastatic cancer patients is critical to improve patient outcomes. Existing pathology and gene expression-based techniques often have limited performance. In this study, we developed a deep neural network (DNN)-based classifier for cancer origin prediction using DNA methylation data of 7,339 patients of 18 different cancer origins from The Cancer Genome Atlas (TCGA). This DNN model was evaluated using four strategies: (1) when evaluated by 10-fold cross-validation, it achieved an overall specificity of 99.72% (95% CI 99.69%-99.75%) and sensitivity of 92.59% (95% CI 91.87%-93.30%); (2) when tested on hold-out testing data of 1,468 patients, the model had an overall specificity of 99.83% and sensitivity of 95.95%; (3) when tested on 143 metastasized cancer patients (12 cancer origins), the model achieved an overall specificity of 99.47% and sensitivity of 95.95%; and (4) when tested on an independent dataset of 581 samples (10 cancer origins), the model achieved overall specificity of 99.91% and sensitivity of 93.43%. Compared to existing pathology and gene expression-based techniques, the DNA methylation-based DNN classifier showed higher performance and had the unique advantage of easy implementation in clinical settings.

Introduction

Identification of cancer origins is routinely performed in clinical practice as site-specific treatments improve patient outcomes [1-4]. While some cancer origins are easy to be determined, others are difficult, especially for metastatic and un-differentiated cancer. Cancer origin determination is typically carried out with immunohistochemistry panels on the tumor specimen and imaging tests, which need considerable resources, time, and expense. In addition, pathologic-based procedures have limited accuracy (66-88%) in determining the origins of metastatic cancer [5-8].

Several gene expression- or microRNA-based molecular classifiers have been developed to identify cancer origin. A k-nearest neighbor classifier based on 92 genes showed an accuracy of 84% in identifying primary site of metastatic cancer via cross-validation [9]. Pathwork, a commercially available platform based on similarity score of 1,550 genes between cancer tissue and reference tissue, achieved an overall sensitivity of 88%, an overall specificity of 99% and an accuracy of 89% in identifying tissue of origin [10, 11]. A decision-tree classifier based on 48 microRNA showed an accuracy of 85-89% in identification of cancer primary sites [12, 13], and an updated version, the 64-microRNA based assay, exhibited an overall sensitivity of 85% [14, 15]. A recent support vector machine-based classifier that integrated gene expression and histopathology showed an accuracy of 88% in known origins of cancer samples [16]. All these molecular platforms have shown better performance in identifying tissue of origin as compared to pathology-based methods. However, gene expression- or microRNA-bases classifiers need to handle RNA that is unstable and less convenient in clinic settings. In addition, these classifiers have performance of <90% accuracy, which may further limit their wide adoption in clinical settings. Hence, it is desirable to develop higher performance prediction tools for cancer origin determination, which can also be easily implemented in clinical settings.

DNA methylation is a process by which methyl groups are added to the DNA molecule and 70-80% of human genome is methylated [17]. It has been shown that DNA methylation is established in tissue specific manner during development [18, 19]. Though the genomes of cancer patients exhibit overall demethylation, tissue specific DNA methylation markers might be conserved [19]. Indeed, a random forest-based cancer origin classifier using DNA methylation was reported to achieve a performance with 88.6% precision and 97.7% recall in the validation set [20], which demonstrated the usefulness of methylation data in cancer origin prediction. Recently, deep learning technologies have rapidly applied to the biomedical field, including protein structure prediction, gene expression regulation, behavior prediction, disease diagnosis and drug development [21, 22]. Studies show that deep learning-based models often achieved higher performance than traditional machine learning methods (e.g. random forest and support vector machine, etc.) in many settings, such as gene expression inference [23], transcript factor binding prediction [24], protein-protein interaction prediction [25], detection of rare disease-associated cell subsets [26], variant calling [27], clinic trial outcome prediction [28], among others. In this study, we trained and robustly evaluated a high-performance cancer origin predictive model by leveraging the large amount of DNA methylation data available in The Cancer Genome Atlas (TCGA) and the recent developments in deep neural network learning techniques. We demonstrated that our model performed better than traditional pathology- or gene expression-based models as well as methylation-based random forest prediction model.

Materials and methods

Datasets

DNA methylation data (Illumina human methylation 450k BeadChip) and clinical information of 8,118 patients across 24 tissue types were obtained from in GDC data portal [29] using TCGAbiolink (Bioconductor package, version 2.5.12) [30]. We excluded six tissue types with less than 100 cases in TCGA to build robust cancer origin classifier. The final data include DNA methylation data and clinical information from 7,339 patients of 18 cancer origins. TCGA data were used for both cancer origin classifier training and evaluation, which were randomly and stratified split into training set (n=4,403), development set (n=1,468) and test set (n=1,468) (Fig 1).

Fig 1. Distribution of cancer samples in TCGA by tissue of origin. A total of 7339 patients were randomly and stratified split into train, dev and test sets according to 60:20:20.

In order to evaluate the classifier trained on TCGA dataset using independent data, we obtained 11 DNA methylation datasets (Illumina 450k platform) from Gene Expression Omnibus (GEO) [31] using GEOquery (Bioconductor package, version 2.42.0) [32]. A total of 581 cancer patients covering 10 cancer origins were obtained and the information for each dataset was described in Table 1.

Table 1. Characteristics of GEO datasets

GEO ID	**Disease**	**Cancer origin**	**Cancer type**	**Num. of patients**
GSE77871	Adrenocortical carcinomas	Adrenal gland	Primary	18
GSE78751	Triple negative breast cancer	Breast	Primary, metastatic	23 12
GSE101764	Colorectal cancer	Colorectal	Primary	112
GSE38268	Head and Neck Squamous Cell Carcinoma	Head and neck	Primary	6
GSE89852	hepatocellular carcinomas	Liver	Primary	37
GSE49149	Pancreatic cancer	Pancreas	Primary	167
GSE112047	Prostate cancer	Prostate	Primary	31
GSE38240	Prostate cancer	Prostate	Primary, metastatic	2 6
GSE73549	Prostate cancer	Prostate	Metastatic	18
GSE86961	Papillary thyroid cancer	Thyroid	Primary	82
GSE52955	Urology cancer	Kidney, Bladder, prostate	Primary	17, 25, 25

Feature selection

Only the training data (n=4,403) from TCGA were used for feature selection. Currently, Illumina 450K and 27K are two commonly used platforms for genome wide analysis of DNA methylation, which measure DNA methylation of around 450K and 27K CpG sites respectively. DNA methylation level of

CpG site is expressed as beta value using the ratio of intensities between methylated and unmethylated alleles. Beta value is between 0 and 1 with 0 being unmethylated and 1 fully methylated. To make the model with good compatibility and also reduce the dimensionality, we firstly reduced CpG sites to 27K for 450K derived samples. To further remove the noise in the data, we used one-way analysis of variance (one-way ANOVA) to filter the CpG sites whose beta values are not significantly different ($p > 0.01$) among different tissues. Then we used the Tukey honest test to remove the CpG sites that maximal differences of their beta values are less than 0.15. The input features used for model building consisted of DNA methylation from 10,360 CpG sites.

Training a deep neural network (DNN) model for cancer origin classification

We used DNA methylation data from training set (n=4,403) to build a DNN model to predict cancer origins. Tensorflow [33], an open source framework to facilitate deep learning model training, was used for this purpose. Four well-established techniques were used to optimize the training process, including weight initialization by Xaiver method [34], Adam optimization [35], learning rate decay and mini-batch training. Xaiver method can efficiently avoid gradient disappearance/explosion that random initialization may bring. Adam, a combination of Stochastic Gradient Descent with momentum descendent [36] and RMSprop [37], makes training process faster. Exponential learning decay (decay every 1,000 steps with a base of 0.96) was used to improve model performance. Training was performed in 128 mini-batch of 30 epochs to efficiently use the data. In addition, three hyperparameters (learning rate, number of hidden layer and hidden layer unit) were optimized to obtain best performance according to development set performance (1,468 patients with the same distribution of cancer origins as training set).

Validating and testing DNN-based cancer origin prediction model

We used four strategies to evaluate the performance of the DNN cancer origin classifier: (1)

evaluation in the10-fold cross-validation in training dataset to obtain overall specificity, sensitivity, PPV and NPV as well as corresponding confidence intervals of this model; (2) evaluation in the hold-out testing dataset to obtain both the overall model performance and tissue-wise performance; (3) evaluation in the subset of metastatic cancer samples nested in testing dataset to assess the performance of the model in predicting the primary sites of metastatic cancer, which are often more difficult to be identified in clinical practice and more clinically relevant; (4) evaluation in independent datasets from GEO to test the robustness and generalizability of this DNN model. Metrics including specificity, sensitivity, positive predictive value (PPV) and negative predictive value (NPV) were reported. Receiver Operating Characteristic curve (ROC curve) was also calculated for each test data performance.

Source code, data availability, and reproducibility

Source code used in this study is publicly available in a Github repository (https://github.com/thunder001/Cancer_origin_prediction). We also shared a Jupyter Notebook to replicate all the machine learning experiments from data processing, model building and optimization to model evaluation. To execute this notebook, the environment needs to be firstly created according to a YAML file available in Github. In addition, we also created a Docker image available in Docker hub (https://hub.docker.com/r/thunder001/cancer_origin_prediction), where you can download it and run the container directly on your computer.

Results

The overall performance of the DNN-based cancer origin classifier in 10-fold cross-validation setting

We used DNA methylation data of 7,339 patients from TCGA across 18 primary tissues to train and test a DNN-based cancer origin classifier. The sample distribution in different cancer origins were shown in Fig 1. The final DNN architecture consists of one input layer (10,360 neurons), two hidden layers (64 neurons each layer) and one output layer (18 neurons) that represents 18 cancer origins (Fig 2).

Figure 2. Schematic representation of DNN architecture of cancer origin classifier.

Evaluated in a 10-fold cross-validation setting, the model achieved an overall precision (positive predictive value, PPV) of 0.9503 (95% CI:0.9373-0.9633) and recall (sensitivity) of 0.9259 (95% CI:0.9187-0.9330) respectively. In addition, this model also achieved a high specificity of 0.9972 (95% CI:0.9969-0.9975) (Table 2).

Table 2. DNN model performance using 10-fold cross validation of training data.

	Mean	SD	CI (95%)
Specificity	0.9972	0.0001	0.9969, 0.9975
Sensitivity (Recall)	0.9259	0.0032	0.9187, 0.9330
PPV (Precision)	0.9503	0.0057	0.9373, 0.9633
NPV	0.9973	0.0001	0.9970, 0.9976

Note: PPV: positive predictive value; NPV: negative predictive value.

DNN-based cancer origin classifier shows high performance in testing dataset

We tested the classifier using test dataset, which includes 1,468 samples with similar distribution with training set (Fig 1). Cancer origin classification and a confusion matrix for all samples were shown in S1 and S2 Tables respectively. Model performance metrics were shown on Table 3. The specificity and negative predictive value (NPV) in individual cancer origin prediction were consistently higher than 0.99. The overall precision (PPV) and recall (sensitivity) reached 0.9608 and 0.9595 respectively. For many cancer tissue origin predictions, including brain, colorectal, prostate, skin, testis, thymus and thyroid, this

DNN model achieved a precision of 100% (Table 3) and an average AUC of 0.99 (Fig 3).

Table 3. DNN model performance in test set.

CANER ORIGIN	SPECIFICITY	SENSITIVITY (RECALL)	PPV (PRECISION)	NPV
AG	0.9993	0.9787	0.9787	0.9993
BLADDER	0.9986	0.9878	0.9759	0.9993
BRAIN	1.0000	1.0000	1.0000	1.0000
BREAST	0.9977	1.0000	0.9810	1.0000
COLORECTAL	1.0000	0.9861	1.0000	0.9993
ESOPHAGUS	0.9909	0.7410	0.7579	0.9902
HN	0.9971	0.9099	0.9619	0.9927
KIDNEY	0.9993	1.0000	0.9925	1.0000
LIVER	0.9993	0.9851	0.9851	0.9993
LUNG	0.9984	0.9740	0.9894	0.9961
PANCREAS	0.9979	1.0000	0.9167	1.0000
PROSTATE	1.0000	1.0000	1.0000	1.0000
SKIN	1.0000	1.0000	1.0000	1.0000
SOFT TISSUE	0.9993	0.9825	0.9825	0.9993
STOMACH	0.9921	0.9375	0.8721	0.9964
TESTIS	1.0000	1.0000	1.0000	1.0000
THYMUS	1.0000	0.8889	1.0000	0.9979
THYROID	1.0000	1.0000	1.0000	1.0000
OVERALL	0.9983	0.9595	0.9608	0.9983

Note: PPV: positive predictive value; NPV: negative predictive value; AG: Adrenal Gland; HN: Head and Neck

Fig 3. AUCs for individual cancer origin prediction in TCGA test set.

There are some variations in precision and recall in different cancer origin predictions. The lowest performance occurred in esophagus origin prediction with a precision of 0.7579 and a recall of 0.7410. A total of 10 of 39 esophagus origins were incorrectly predicted as stomach origins (S1 and S2 Tables). Given that esophagus is a broad area, if a tumor is located at the border of stomach and esophagus, it might be difficult for the classifier to distinguish these two tissues. In addition, tissues from adjacent regions may have similar methylation profiles so that the methylation-based prediction model has difficulty in differentiating cancers with adjacent origins (e.g., esophagus vs stomach).

DNN-based cancer tissue classifier shows high performance in determining the origins of metastasized cancers

We evaluated the performance of the classifier in determining the origins of metastatic cancers that nested in our test data. Our data contained 701 samples from distantly metastasized cancers and 558 of them have been used for model development. We then used remaining 143 samples from 12 cancer origins with various sample sizes for evaluation (Fig 4A). Cancer origin predictions and corresponding confusion matrix were shown in S3 and S4 Tables. Model performance metrics and ROC curves were shown in Table 4 and Fig 4B. Consistently, DNN model showed robust high performance in predicting metastatic cancer origins.

Fig 4. Performance of the DNN-based cancer origin classifier in metastatic cancer samples from TCGA test set. (A) Distribution of metastatic cancer samples by tissue of origin. (B) AUCs for individual cancer origin prediction

Table 4. DNN model performance in metastatic cancer samples.

CANER ORIGIN	SPECIFICITY	SENSITIVITY	PPV	NPV
		(RECALL)	(PRECISION)	
ADRENAL GLAND	1.0000	1.0000	1.0000	1.0000
BLADDER	1.0000	0.9643	1.0000	0.9914
BREAST	0.9929	1.0000	0.7500	1.0000
COLORECTAL	1.0000	1.0000	1.0000	1.0000
ESOPHAGUS	0.9504	1.0000	0.2222	1.0000
HEAD AND NECK	1.0000	0.8833	1.0000	0.9222
KIDNEY	1.0000	1.0000	1.0000	1.0000
LIVER	0.9929	1.0000	0.6667	1.0000
LUNG	1.0000	0.6667	1.0000	0.9929
PANCREAS	1.0000	1.0000	1.0000	1.0000
STOMACH	1.0000	1.0000	1.0000	1.0000
THYROID	1.0000	1.0000	1.0000	1.0000
OVERALL	0.9947	0.9595	0.8866	0.9922

Note: PPV: positive predictive value; NPV: negative predictive value.

We noticed that performance metrics in several cancer origin predictions were poor: a precision of 0.22 for esophagus origin prediction, a precision of 0.67 for liver origin prediction and a recall of 0.67 for lung prediction. The poor performance in these three cancer origin predictions may be due to small sample size. As mentioned above, metastatic cancer samples comprise only a small subset of test dataset in TCGA, the majority of which are primary tumors. Only 2, 2 and 3 metastatic cancer samples from esophagus, liver and lung origin respectively were included in test dataset (Fig 4A). The classifier misclassified 6 out of 60 head and neck cancers as esophagus origin and 1 of 3 of lung cancers as liver cancers (S4 Table). Due to small sample sizes for esophagus, liver and lung cancers, a few misclassifications had significant impacts on the precision metrics.

DNN-based cancer tissue classifier shows high performance in independent testing datasets

The DNN model was trained using DNA methylation data from TCGA. We then tested it in independent datasets of 11 data series consisting of 581 tumor samples covering 10 tissue origins downloaded from Gene Expression Omnibus (GEO). The sample distribution was shown in Fig 5A and cancer origin predictions were listed in S5 Table. Evaluated using these independent datasets, the DNN model achieved high performance with an overall precision and recall of 98.69% and 93.43% respectively (Table 5). High performance was also achieved in individual cancer origin predictions (Table 5) with an average AUC of 0.99 (Fig 5B). Importantly, the model achieved 100% accuracy in predicting the origins of metastatic cancers in these datasets, including 24 prostate cancer that metastasized to bone, lymph node or soft tissue and 12 breast cancer that metastasized to lymph node (see Table 1 for these samples).

Fig 5. Performance of the DNN-based cancer origin classifier in GEO dataset. (A) Distribution of cancer samples obtained from GEO by tissue of origin. (B) AUCs for individual cancer origin prediction

Table 5. DNN model performance using independent cancer samples (GEO)

CANER	SPECIFICITY	SENSITIVITY	PPV	NPV
ORIGIN		(RECALL)	(PRECISION)	
ADRENAL GLAND	1.0000	0.7778	1.0000	0.9929
BLADDER	1.0000	1.0000	1.0000	1.0000
BREAST	0.9963	0.9714	0.9444	0.9982
COLORECTAL	1.0000	0.9643	1.0000	0.9915
HEAD AND NECK	1.0000	0.8333	1.0000	0.9983
KIDNEY	1.0000	1.0000	1.0000	1.0000
LIVER	0.9945	1.0000	0.9250	1.0000
PANCREAS	1.0000	0.8084	1.0000	0.9283
PROSTATE	1.0000	1.0000	1.0000	1.0000
THYROID	1.0000	0.9878	1.0000	0.9980
OVERALL	0.9991	0.9343	0.9869	0.9907

Note: PPV: positive predictive value; NPV: negative predictive value.

Discussion

We developed a deep neural network model to predict the cancer origins based on large amount of DNA methylation data from 7,339 patients of 18 different cancer origins. By combining DNA methylation data with deep learning algorithm, our caner origin classifier achieved high performance as demonstrated in four different evaluation settings. Compared with Pathwork, a commercially available cancer origin classifier based on gene expressions [10], our DNN model showed higher precision (95.03% vs 89.4%) and recall (92.3% vs 87.8%) and comparable specificity (99.7% vs 99.4%). Compared with DNA methylation-based random forest model, our DNN model achieved higher PPV (precision) (95.03% in cross validation and 96.08% in test vs 88.6%) and comparable specificity, sensitivity and NPV. In addition, we showed that our DNN model is highly robust and generalizable as evaluated in an independent testing dataset of 581 samples (10 cancer origins), with overall specificity of 99.91% and sensitivity of 93.43%. Therefore, high performance both in primary and metastatic cancer origin prediction and the potential for easy implementation in clinical setting make the methylation-based DNN model a promising tool in determining cancer origins.

DNA methylation is established in tissue specific manner and conserved during cancer development [19], which makes DNA methylation profile a very useful feature in cancer origin prediction. Deep neural networks (DNNs) excels in capturing hierarchical features inherent in many complicated biological mechanisms. Our study indicates that the trained DNN model may be able to capture hierarchical patterns of cancer origins from the DNA methylation data. While Interpretation of deep learning-based models is a rapidly developing field and we expect that our model can be explained in a meaningful way in the future.

Our DNN model has potential in predicting origins of Cancer of Unknown Primary origin (CUP). CUP is a sub-group of heterogenous metastatic cancer with illusive primary site even after standard pathological examination [38]. It is estimated that 3-5% metastatic cancers are CUP and the majority of CUP patients (80%) have poor prognosis with overall survival of 6 -10 months [38]. Identifying primary site of CUP poses challenges for treatment decisions in clinical practice. Currently, intensive pathologic examination still leaves 30% of them unidentified [39, 40]. High performance of our DNA methylation-based DNN model may provide an opportunity in this scenario when pathology-based approach fails. However, due to the limited CUP data in both TCGA and GEO, we currently are unable to test the DNN models in predicting the origins of CUP. Our future direction is to collaborate with hospital to collect DNA methylation data from CUP patients to test our model. One challenge is to obtain the true primary sites for these patients. Due to unknown property of CUP, true primary sites may be established in later cancer development [20]. Another is through the post-mortem examination of patients since 75% of primary sites of CUP were found in autopsy [41].

One limitation of this study is that small sizes of metastatic cancers in our data. Two resources of metastatic cancer were used in this study: TCGA and GEO. TCGA has 701 metastatic cancer samples (12 tissues) with available methylation data from Illumina Human Methylation 450K platform. While the model achieved an overall specificity of 99.47% and sensitivity of 95.95% in cross-validation using TCGA data, we were unable to robustly test it using independent dataset since methylation data of metastatic cancers is limited in GEO. Further independent validation of our DNN-based model in predicting origins of metastatic cancers, especially poorly differentiated or undifferentiated metastatic cancer samples, is needed.

References

1. Hainsworth JD, Rubin MS, Spigel DR, Boccia RV, Raby S, Quinn R, et al. Molecular gene expression profiling to predict the tissue of origin and direct site-specific therapy in patients with carcinoma of unknown primary site: a prospective trial of the Sarah Cannon research institute. J Clin Oncol 2013. 10;31:217-23.
2. Varadhachary GR, Raber MN, Matamoros A, Abbruzzese JL. Carcinoma of unknown primary with a colon-cancer profile-changing paradigm and emerging definitions. Lancet Oncol. 2008;9:596–9.
3. Varadhachary GR, Spector Y, Abbruzzese JL, Rosenwald S, Wang H, Aharonov R, et al. Prospective gene signature study using microRNA to identify the tissue of origin in patients with carcinoma of unknown primary. Clin Cancer Res. 2011;17:4063-70.
4. Varadhachary GR, Karanth S, Qiao W, Carlson HR, Raber MN, Hainsworth JD, et al. Carcinoma of unknown primary with gastrointestinal profile: immunohistochemistry and survival data for this favorable subset. Int J Clin Oncol. 2014;19:479-84.
5. Brown RW, Campagna LB, Dunn JK, Cagle PT. Immunohistochemical identification of tumor markers in metastatic adenocarcinoma. A diagnostic adjunct in the determination of primary site. Am J Clin Pathol. 1997;107:12-9.
6. DeYoung BR, Wick MR. Immunohistologic evaluation of metastatic carcinomas of unknown origin: an algorithmic approach. Semin Diagn Pathol. 2000;17:184-93.
7. Dennis JL, Hvidsten TR, Wit EC, Komorowski J, Bell AK, Downie I, et al. Markers of adenocarcinoma characteristic of the site of origin: development of a diagnostic algorithm. Clin Cancer Res. 2005;11:3766-72.
8. Park SY, Kim BH, Kim JH, Lee S, Kang GH. Panels of immunohistochemical markers help

determine primary sites of metastatic adenocarcinoma. Arch Pathol Lab Med. 2007;131:1561-7

9. Ma XJ, Patel R, Wang X, Salunga R, Murage J, Desai R, et al. Molecular classification of human cancers using a 92-gene real-time quantitative polymerase chain reaction assay. Arch Pathol Lab Med. 2006;130:465–73.

10. Monzon FA, Lyons-Weiler M, Buturovic LJ, Rigl CT, Henner WD, Sciulli C, et al. Multicenter validation of a 1,550-gene expression profile for identification of tumor tissue of origin. J Clin Oncol. 2009;27:2503–8.

11. Pillai R, Deeter R, Rigl CT, Nystrom JS, Miller MH, Buturovic L, et al. Validation and reproducibility of a microarray-based gene expression test for tumor identification in formalin-fixed, paraffin-embedded specimens. J Mol Diagn. 2011;13:48-56.

12. Rosenfeld N, Aharonov R, Meiri E, Rosenwald S, Spector Y, Zepeniuk M, et al. MicroRNAs accurately identify cancer tissue origin. Nat Biotechnol. 2008;26:462–9.

13. Rosenwald S, Gilad S, Benjamin S, Lebanony D, Dromi N, Faerman A, et al. Validation of a microRNA-based qRT-PCR test for accurate identification of tumor tissue origin. Mod Pathol 2010;23:814–23.

14. Meiri E, Mueller WC, Rosenwald S, Zepeniuk M, Klinke E, Edmonston TB, et al. A second-generation microRNA-based assay for diagnosing tumor tissue origin. Oncologist. 2012;17:801–12

15. Pentheroudakis G, Pavlidis N, Fountzilas G, Krikelis D, Goussia A, Stoyianni A, et al. Novel microRNA-based assay demonstrates 92% agreement with diagnosis based on clinicopathologic and management data in a cohort of patients with carcinoma of unknown primary. Mol Cancer. 2013;12:57.

16. Tothill RW, Shi F, Paiman L, Bedo J, Kowalczyk A, Mileshkin L, et al. Development and validation of a gene expression tumour classifier for cancer of unknown primary. Pathology. 2015;47:7–12.

17. Kulis M, Esteller M. DNA methylation and cancer. Adv Genet. 2010;70:27–56.

18. Ohgane J, Yagi S, Shiota K. Epigenetics: the DNA methylation profile of tissue-dependent and differentially methylated regions in cells. Placenta. 2008;29 Suppl A:S29–35.

19. Fernandez AF, Assenov Y, Martin-Subero JI, Balint B, Siebert R, Taniguchi H, et al. A DNA methylation fingerprint of 1628 human samples. Genome Res. A DNA methylation fingerprint of 1628 human samples. Genome Res. 2012;22:407–19.

20. Moran S, Martínez-Cardús A, Sayols S, Musulén E, Balañá C, Estival-Gonzalez A, et al. Epigenetic profiling to classify cancer of unknown primary: a multicentre, retrospective analysis. Lancet Oncol. 2016;17:1386–1395.

21. Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2017;18:851–869

22. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141). doi: 10.1098/rsif.2017.0387

23. Chen Y, Li Y, Narayan R, Subramanian A, Xie X. Gene expression inference with deep learning. Bioinformatics. 2016;32(12):1832-9.

24. Alipanahi B, Delong A, Weirauch MT, Frey BJ. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat Biotechnol. 2015;33(8):831-8

25. Du T, Liao L, Wu CH, Sun B. Prediction of residue-residue contact matrix for protein-protein interaction with Fisher score features and deep learning. Methods. 2016;110:97-105

26. Arvaniti E, Claassen M. Sensitive detection of rare disease-associated cell subsets via representation learning. Nat Commun. 2017;8:14825. doi: 10.1038/ncomms14825.

27. Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983-987

28. Artemov AV, Putin E, Vanhaelen Q, Aliper A, Ozerov IV, Zhavoronkov A, et al. Integrated deep learned transcriptomic and structure-based predictor of clinical trials outcomes. BioRxiv [Preprint]. 2016. (doi:10.1101/095653)

29. GDC data portal. https://portal.gdc.cancer.gov. Accessed 7 August 2019

30. Colaprico A, Silva TC, Olsen C, Garofano L, Cava C, Garolini D, et al. TCGAbiolinks: a R/Bioconductor package for integrative analysis of TCGA data. Nucleic Acids Res. 2016;44:e71

31. Gene Expression Omnibus. https://www.ncbi.nlm.nih.gov/geo/. Accessed 7 August 2019
32. Davis S, Meltzer PS. GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor. Bioinformatics. 2007;23:1846–7
33. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-scale machine learning on heterogeneous systems. In: OSDI'16 Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation. 2016;265-283
34. Glorot X, Bengio Y. Understanding the difficulty of training deep feedforward neural networks. In: Proceedings of the International Conference on Artificial Intelligence and Statistics. 2010;249-256
35. Diederik P. Kingma and Jimmy Lei Ba. Adam. A method for stochastic optimization. arXiv. 2014;1412.6980v9
36. Qian N. On the momentum term in gradient descent learning algorithms. Neural Netw. 1999;12:145-151.
37. Mcmahan HB and Streeter M. Delay-Tolerant Algorithms for Asynchronous Distributed Online Learning. Advances in Neural Information Processing Systems (Proceedings of NIPS). 2014;1–9.
38. Varadhachary GR, Raber MN. Cancer of unknown primary site. N Engl J Med. 2014;371:757–65
39. Krämer A, Hübner G, Schneeweiss A, Folprecht G, Neben K. Carcinoma of Unknown Primary - an Orphan Disease? Breast Care (Basel). 2008;3:164-170.
40. Ettinger DS, Agulnik M, Cates JM, Cristea M, Denlinger CS, Eaton KD, et al. NCCN Clinical Practice Guidelines Occult primary. J Natl Compr Canc Netw. 2011;9:1358-95.
41. Pentheroudakis G, Golfinopoulos V, Pavlidis N. Switching benchmarks in cancer of unknown primary: from autopsy to microarray. Eur J Cancer. 2007;43:2026–36

Supporting information

S1 Table. Cancer origin predictions for 1468 patient samples from TCGA. (DOCX)

S2 Table. Confusion matrix for TCGA test set predictions. (CSV)

S3 Table. Cancer tissue origin predictions for 143 metastatic cancer samples. (DOCX)

S4 Table. Confusion matrix for metastatic cancer samples in TCGA test set. (CSV)

S5 Table. Cancer origin predictions for 581 samples from GEO datasets. (DOCX)

S6 Table. Confusion matrix for GEO sample predictions. (CSV)

Figure 1

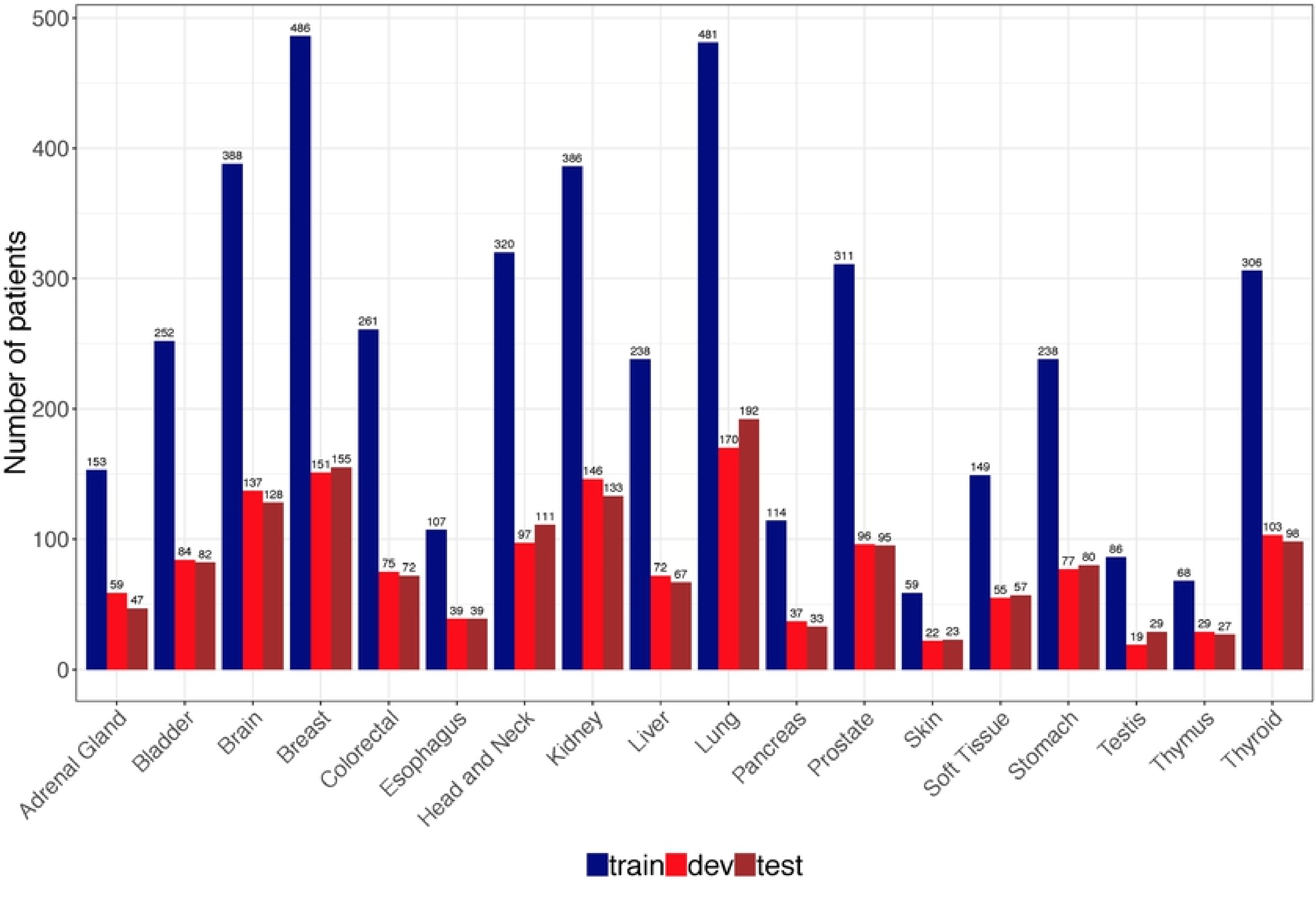
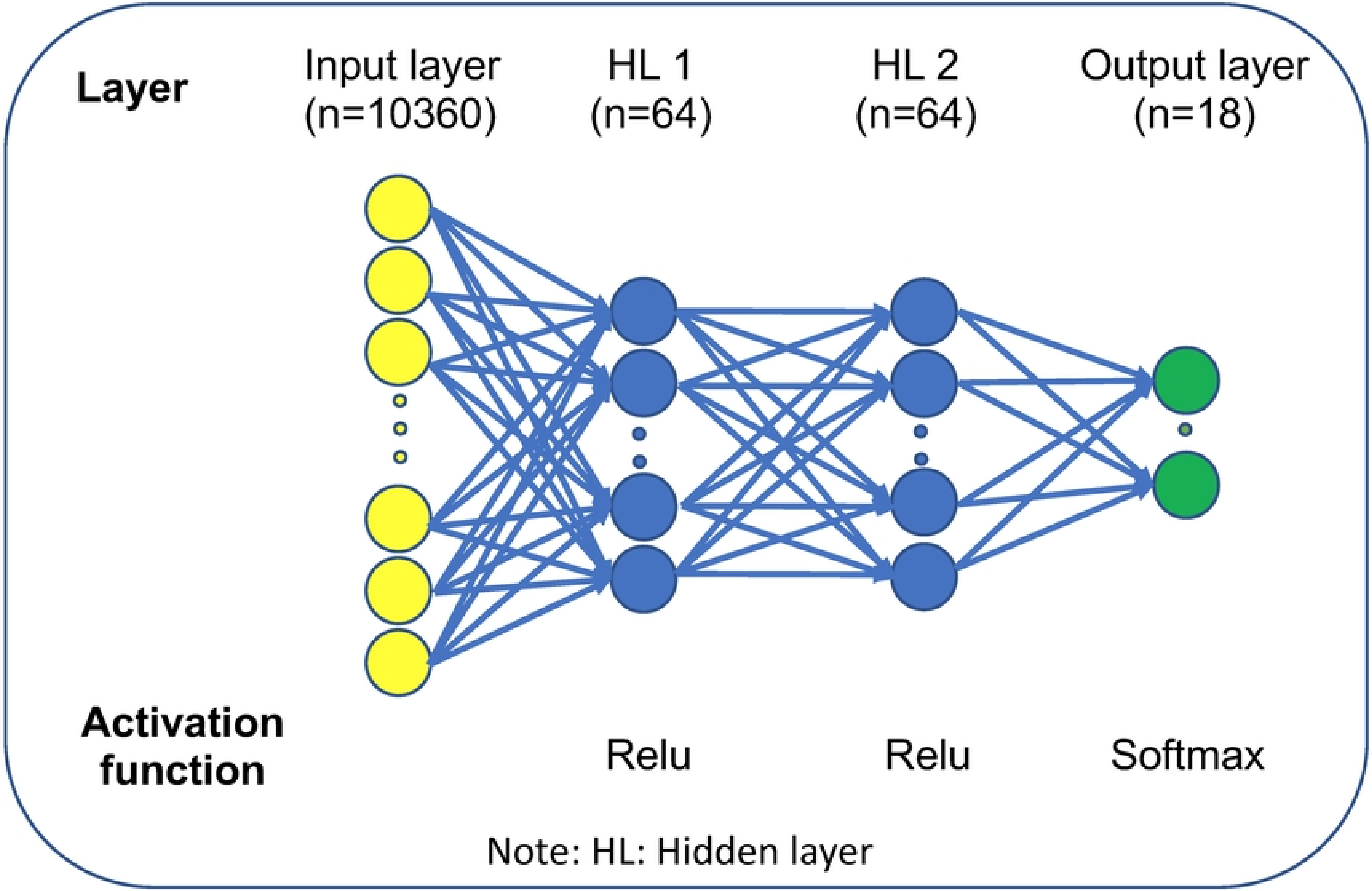

Number of patients
500
400
300
200
100
0
153
59
47
252
84
82
388
137
128
486
151
155
261
75
72
107
39
39
320
97
111
386
146
133
238
72
67
481
170
192
114
37
33
311
96
95
59
22
23
149
55
57
238
77
80
86
19
29
68
29
27
306
103
98
Adrenal Gland
Bladder
Brain
Breast
Colorectal
Esophagus
Head and Neck
Kidney
Liver
Lung
Pancreas
Prostate
Skin
Soft Tissue
Stomach
Testis
Thymus
Thyroid
train
dev
test

Figure 2



Note: HL: Hidden layer

Figure 3

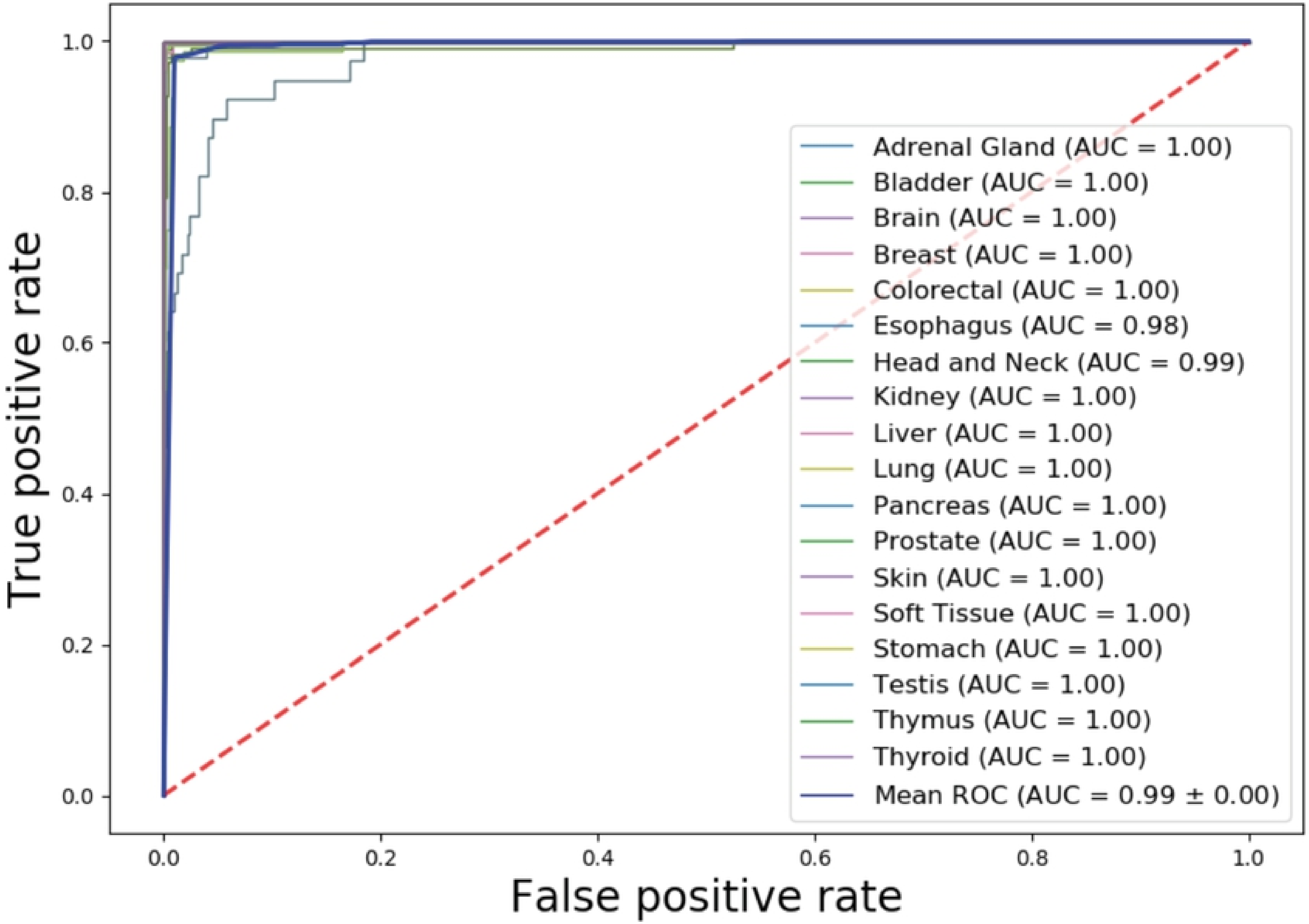

True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Adrenal Gland (AUC = 1.00)
Bladder (AUC = 1.00)
Brain (AUC = 1.00)
Breast (AUC = 1.00)
Colorectal (AUC = 1.00)
Esophagus (AUC = 0.98)
Head and Neck (AUC = 0.99)
Kidney (AUC = 1.00)
Liver (AUC = 1.00)
Lung (AUC = 1.00)
Pancreas (AUC = 1.00)
Prostate (AUC = 1.00)
Skin (AUC = 1.00)
Soft Tissue (AUC = 1.00)
Stomach (AUC = 1.00)
Testis (AUC = 1.00)
Thymus (AUC = 1.00)
Thyroid (AUC = 1.00)
Mean ROC (AUC = 0.99 ± 0.00)

Figure 4

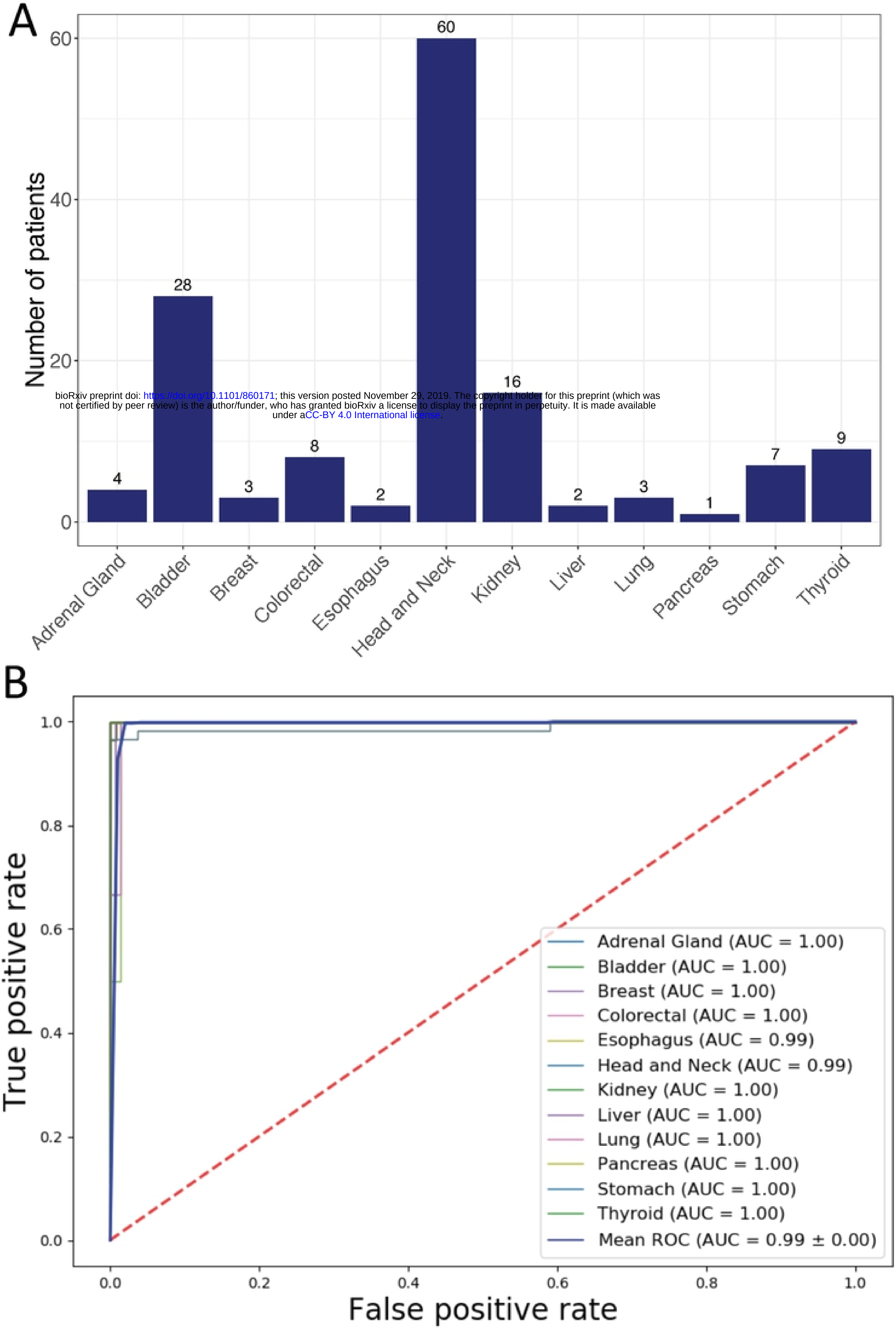
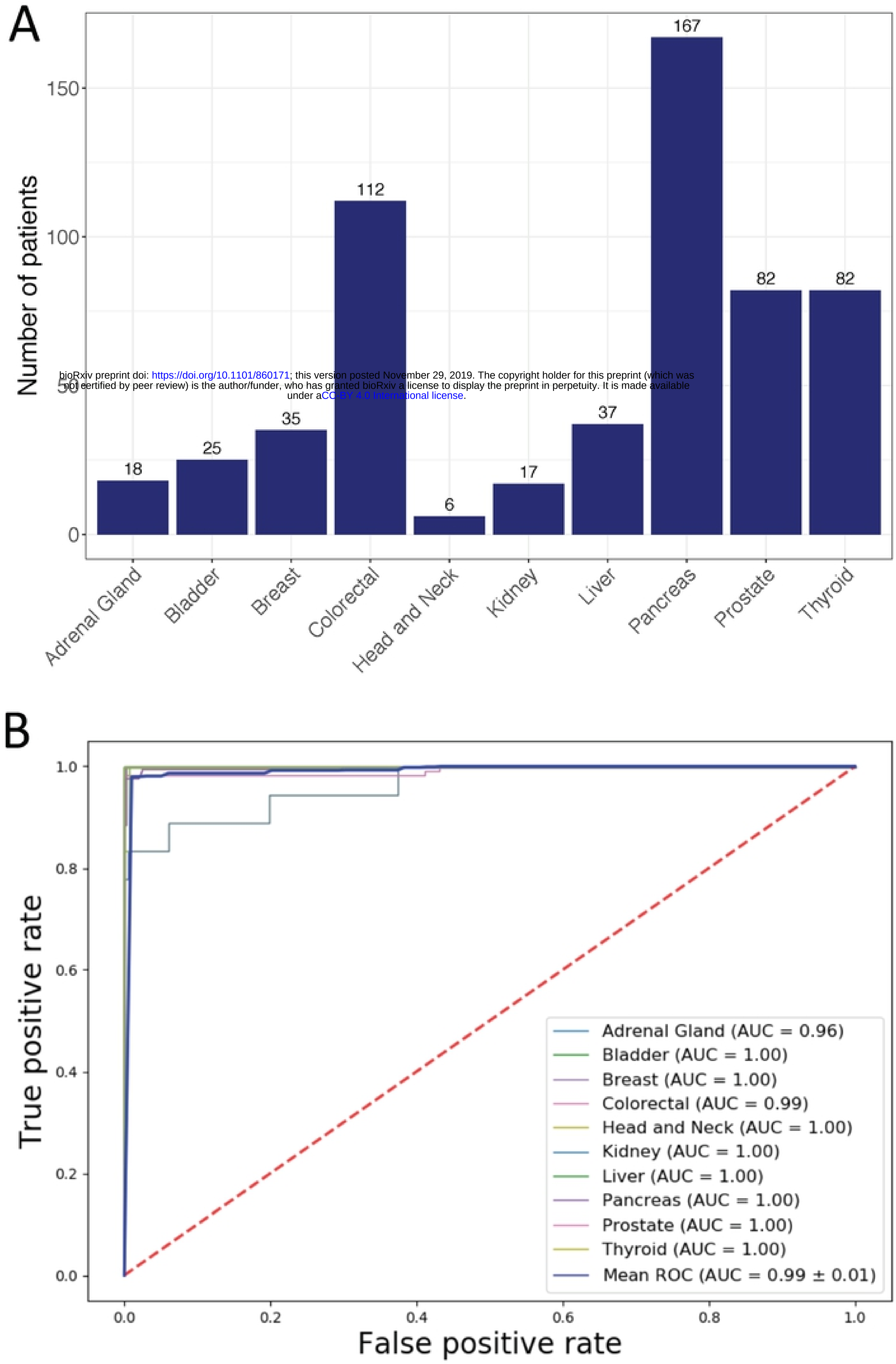

A
Number of patients
0
20
40
60
4
28
3
8
2
60
16
2
3
1
7
9
Adrenal Gland
Bladder
Breast
Colorectal
Esophagus
Head and Neck
Kidney
Liver
Lung
Pancreas
Stomach
Thyroid
B
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Adrenal Gland (AUC = 1.00)
Bladder (AUC = 1.00)
Breast (AUC = 1.00)
Colorectal (AUC = 1.00)
Esophagus (AUC = 0.99)
Head and Neck (AUC = 0.99)
Kidney (AUC = 1.00)
Liver (AUC = 1.00)
Lung (AUC = 1.00)
Pancreas (AUC = 1.00)
Stomach (AUC = 1.00)
Thyroid (AUC = 1.00)
Mean ROC (AUC = 0.99 ± 0.00)

Figure 5


A
Number of patients
0
50
100
150
18
25
35
112
6
17
37
167
82
82
Adrenal Gland
Bladder
Breast
Colorectal
Head and Neck
Kidney
Liver
Pancreas
Prostate
Thyroid
B
True positive rate
0.0
0.2
0.4
0.6
0.8
1.0
False positive rate
0.0
0.2
0.4
0.6
0.8
1.0
Adrenal Gland (AUC = 0.96)
Bladder (AUC = 1.00)
Breast (AUC = 1.00)
Colorectal (AUC = 0.99)
Head and Neck (AUC = 1.00)
Kidney (AUC = 1.00)
Liver (AUC = 1.00)
Pancreas (AUC = 1.00)
Prostate (AUC = 1.00)
Thyroid (AUC = 1.00)
Mean ROC (AUC = 0.99 ± 0.01)