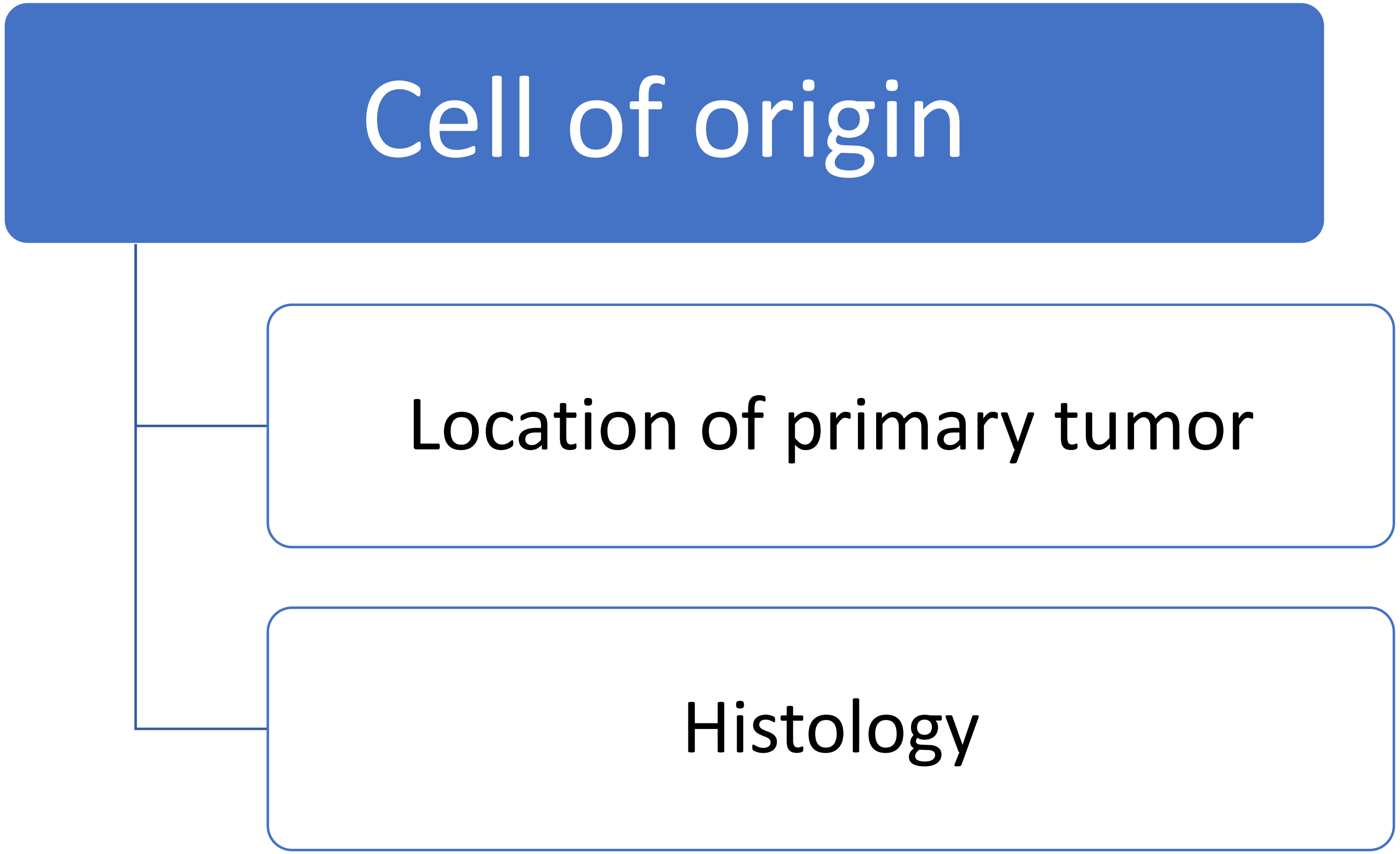
Identifying tumor’s organ of origin using deep learning

Mohammed El-Kebir

Major determinant of outcome



Cell of origin

Location of primary tumor

Histology

Datasets

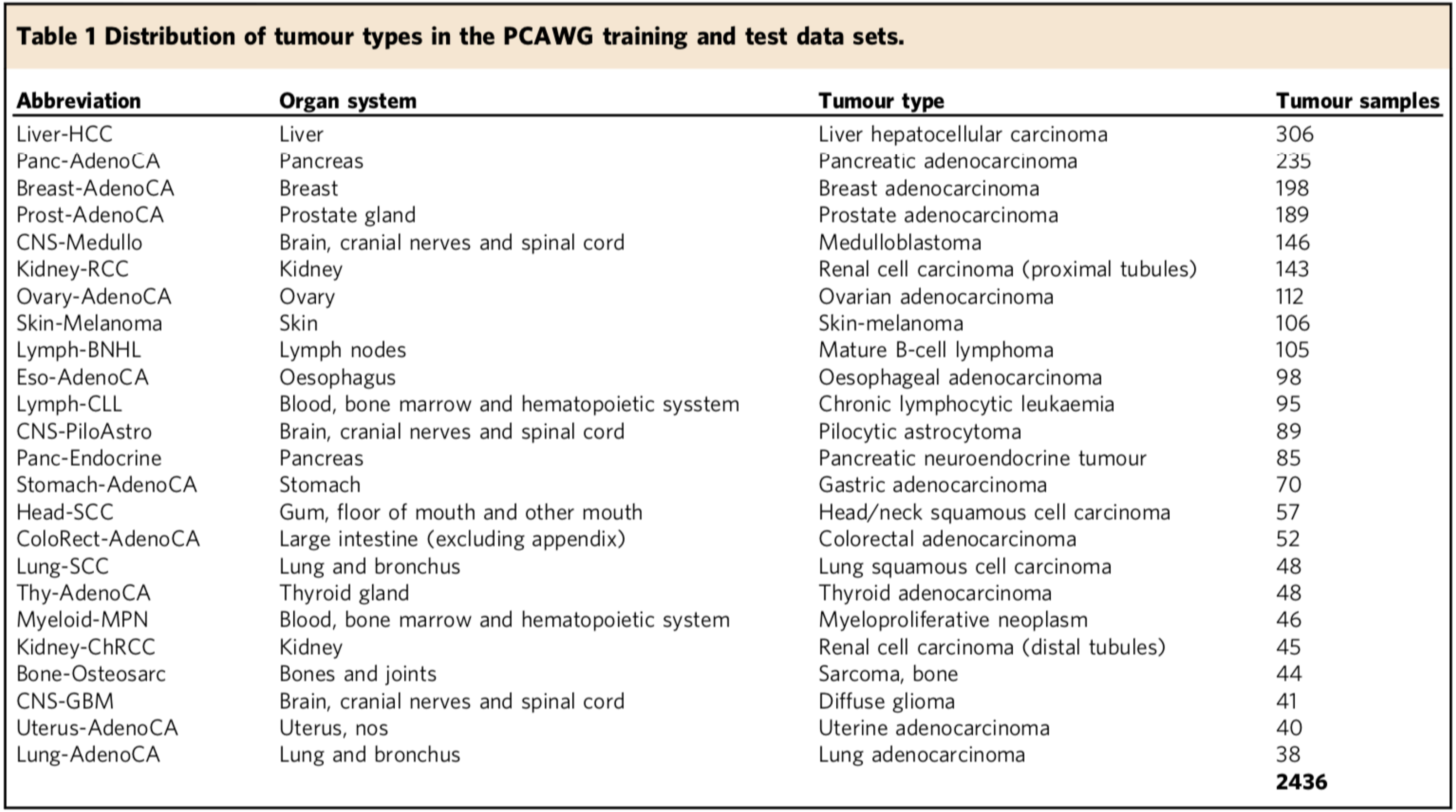


Table 1 Distribution of tumour types in the PCAWG training and test data sets.

Abbreviation	Organ system	Tumour type	Tumour samples
Liver-HCC	Liver	Liver hepatocellular carcinoma	306
Panc-AdenoCA	Pancreas	Pancreatic adenocarcinoma	235
Breast-AdenoCA	Breast	Breast adenocarcinoma	198
Prost-AdenoCA	Prostate gland	Prostate adenocarcinoma	189
CNS-Medullo	Brain, cranial nerves and spinal cord	Medulloblastoma	146
Kidney-RCC	Kidney	Renal cell carcinoma (proximal tubules)	143
Ovary-AdenoCA	Ovary	Ovarian adenocarcinoma	112
Skin-Melanoma	Skin	Skin-melanoma	106
Lymph-BNHL	Lymph nodes	Mature B-cell lymphoma	105
Eso-AdenoCA	Oesophagus	Oesophageal adenocarcinoma	98
Lymph-CLL	Blood, bone marrow and hematopoietic sysstem	Chronic lymphocytic leukaemia	95
CNS-PiloAstro	Brain, cranial nerves and spinal cord	Pilocytic astrocytoma	89
Panc-Endocrine	Pancreas	Pancreatic neuroendocrine tumour	85
Stomach-AdenoCA	Stomach	Gastric adenocarcinoma	70
Head-SCC	Gum, floor of mouth and other mouth	Head/neck squamous cell carcinoma	57
ColoRect-AdenoCA	Large intestine (excluding appendix)	Colorectal adenocarcinoma	52
Lung-SCC	Lung and bronchus	Lung squamous cell carcinoma	48
Thy-AdenoCA	Thyroid gland	Thyroid adenocarcinoma	48
Myeloid-MPN	Blood, bone marrow and hematopoietic system	Myeloproliferative neoplasm	46
Kidney-ChRCC	Kidney	Renal cell carcinoma (distal tubules)	45
Bone-Osteosarc	Bones and joints	Sarcoma, bone	44
CNS-GBM	Brain, cranial nerves and spinal cord	Diffuse glioma	41
Uterus-AdenoCA	Uterus, nos	Uterine adenocarcinoma	40
Lung-AdenoCA	Lung and bronchus	Lung adenocarcinoma	38
			2436

Feature types

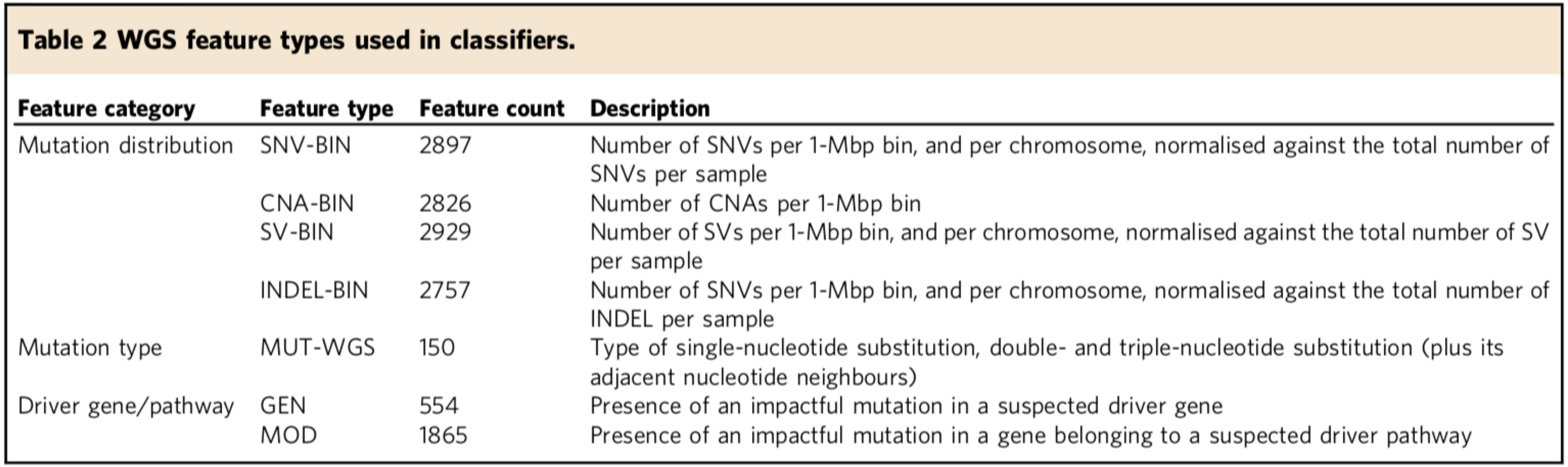
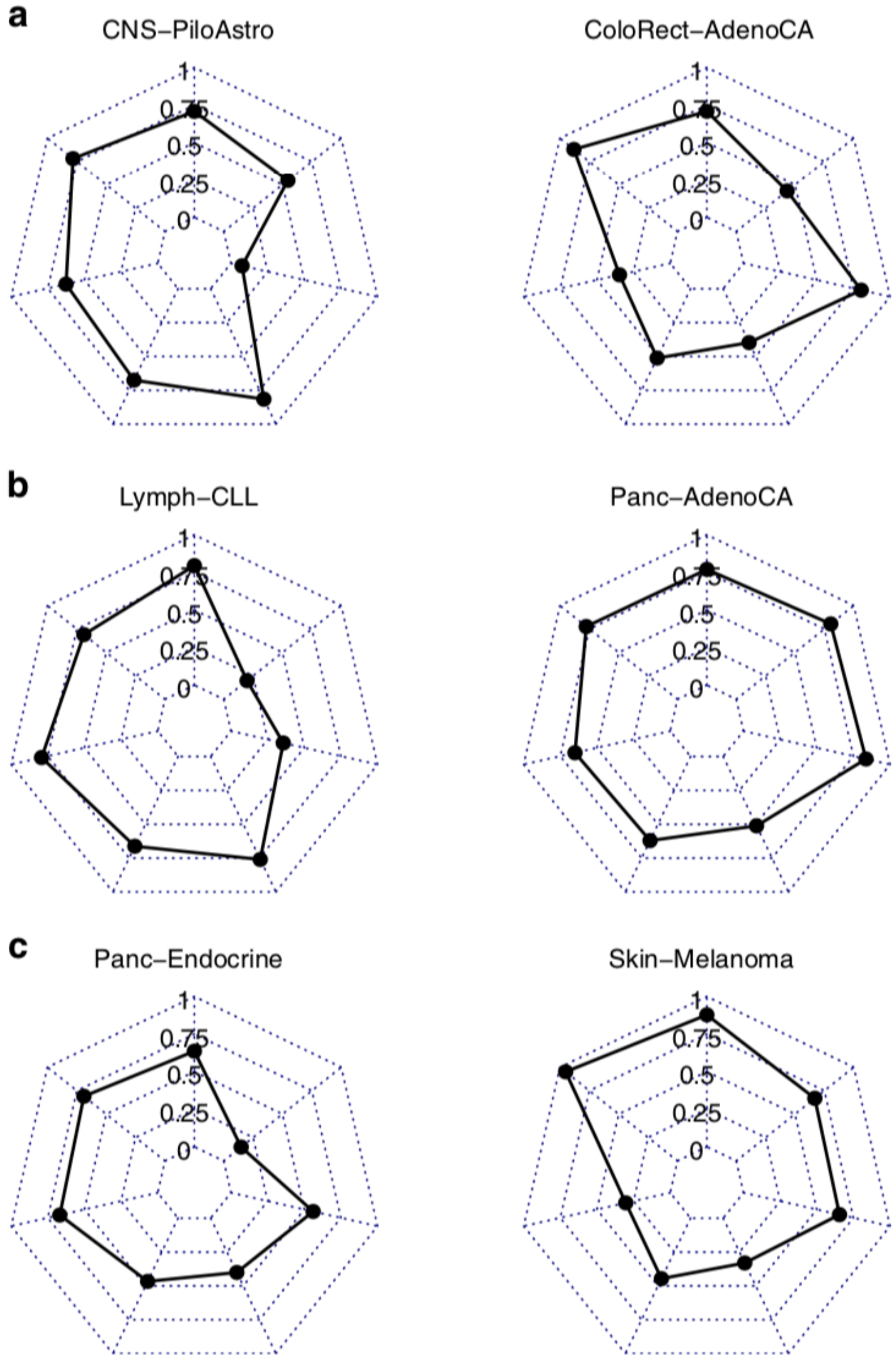


Table 2 WGS feature types used in classifiers.

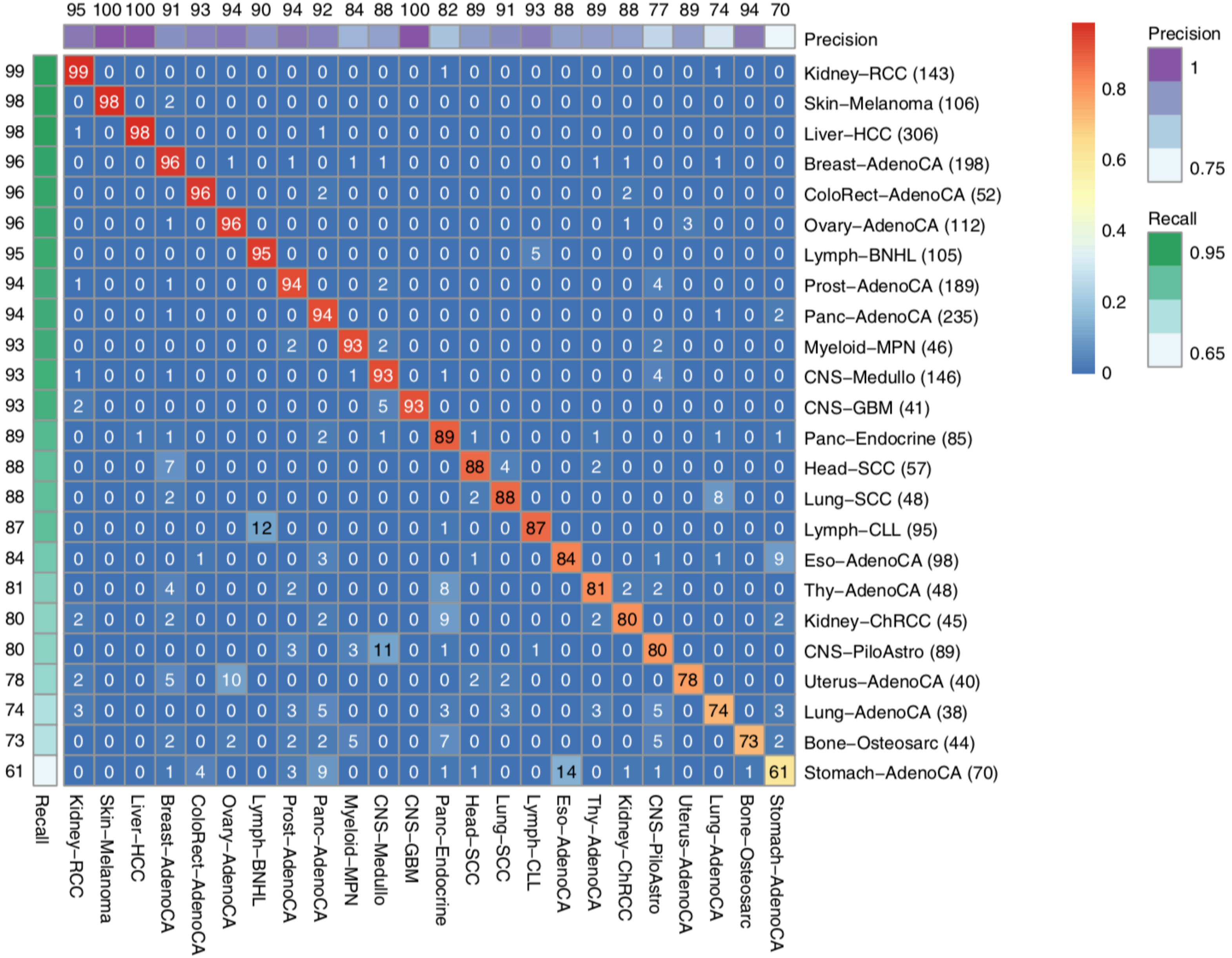
Feature category	Feature type	Feature count	Description
Mutation distribution	SNV-BIN	2897	Number of SNVs per 1-Mbp bin, and per chromosome, normalised against the total number of SNVs per sample
	CNA-BIN	2826	Number of CNAs per 1-Mbp bin
	SV-BIN	2929	Number of SVs per 1-Mbp bin, and per chromosome, normalised against the total number of SV per sample
	INDEL-BIN	2757	Number of SNVs per 1-Mbp bin, and per chromosome, normalised against the total number of INDEL per sample
Mutation type	MUT-WGS	150	Type of single-nucleotide substitution, double- and triple-nucleotide substitution (plus its adjacent nucleotide neighbours)
Driver gene/pathway	GEN	554	Presence of an impactful mutation in a suspected driver gene
	MOD	1865	Presence of an impactful mutation in a gene belonging to a suspected driver pathway

RF for individual features

Fig. 1 Comparison of tumour-type classifiers using single and multiple feature types. a Radar plots describing the cross-validation-derived accuracy (F1) score of Random Forest classifiers trained on each of 7 individual feature categories, across six representative tumour types. **b** Summary of Random Forest classifier accuracy (F1) trained on individual feature categories across all 24 tumour types. **c** Accuracy of classifiers trained on multiple feature categories. *RF Best Models* corresponds to the cross-validation F1 scores of Random Forest classifiers trained on the three best single-feature categories for all 24 tumour types. *DNN Model* shows the distribution of F1 scores for held-out samples for a multi-class neural network trained using passenger mutation distribution and type. *DNN Model + Drivers* shows F1 scores for the neural net when driver genes and pathways are added to the training features. The centre line in the boxplot represents the median of the F1 scores. The lower and upper bounds of the box represent the first and third quartile. The whiskers extend to 1.5 IQR plus the third quartile or minus the first quantile.



Results (1/4)



Recall		Kidney–RCC	Skin–Melanoma	Liver–HCC	Breast–AdenoCA	ColoRect–AdenoCA	Ovary–AdenoCA	Lymph–BNHL	Prost–AdenoCA	Panc–AdenoCA	Myeloid–MPN	CNS–Medullo	CNS–GBM	Panc–Endocrine	Head–SCC	Lung–SCC	Lymph–CLL	Eso–AdenoCA	Thy–AdenoCA	Kidney–ChRCC	CNS–PiloAstro	Uterus–AdenoCA	Lung–AdenoCA	Bone–Osteosarc	Stomach–AdenoCA
	Precision	95	100	100	91	93	94	90	94	92	84	88	100	82	89	91	93	88	89	88	77	89	74	94	70
99	Kidney–RCC (143)	99	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0
98	Skin–Melanoma (106)	0	98	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
98	Liver–HCC (306)	1	0	98	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
96	Breast–AdenoCA (198)	0	0	0	96	0	1	0	1	0	1	1	0	0	0	0	0	0	1	1	0	0	1	0	0
96	ColoRect–AdenoCA (52)	0	0	0	0	96	0	0	0	2	0	0	0	0	0	0	0	0	0	2	0	0	0	0	0
96	Ovary–AdenoCA (112)	0	0	0	1	0	96	0	0	0	0	0	0	0	0	0	0	0	0	1	0	3	0	0	0
95	Lymph–BNHL (105)	0	0	0	0	0	0	95	0	0	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0
94	Prost–AdenoCA (189)	1	0	0	1	0	0	0	94	0	0	2	0	0	0	0	0	0	0	0	4	0	0	0	0
94	Panc–AdenoCA (235)	0	0	0	1	0	0	0	0	94	0	0	0	0	0	0	0	0	0	0	0	0	1	0	2
93	Myeloid–MPN (46)	0	0	0	0	0	0	0	2	0	93	2	0	0	0	0	0	0	0	0	2	0	0	0	0
93	CNS–Medullo (146)	1	0	0	1	0	0	0	0	0	1	93	0	1	0	0	0	0	0	0	4	0	0	0	0
93	CNS–GBM (41)	2	0	0	0	0	0	0	0	0	0	5	93	0	0	0	0	0	0	0	0	0	0	0	0
89	Panc–Endocrine (85)	0	0	1	1	0	0	0	0	2	0	1	0	89	1	0	0	0	1	0	0	0	1	0	1
88	Head–SCC (57)	0	0	0	7	0	0	0	0	0	0	0	0	0	88	4	0	0	2	0	0	0	0	0	0
88	Lung–SCC (48)	0	0	0	2	0	0	0	0	0	0	0	0	0	2	88	0	0	0	0	0	0	8	0	0
87	Lymph–CLL (95)	0	0	0	0	0	0	12	0	0	0	0	0	1	0	0	87	0	0	0	0	0	0	0	0
84	Eso–AdenoCA (98)	0	0	0	0	1	0	0	0	3	0	0	0	0	1	0	0	84	0	0	1	0	1	0	9
81	Thy–AdenoCA (48)	0	0	0	4	0	0	0	2	0	0	0	0	8	0	0	0	0	81	2	2	0	0	0	0
80	Kidney–ChRCC (45)	2	0	0	2	0	0	0	0	2	0	0	0	9	0	0	0	0	2	80	0	0	0	0	2
80	CNS–PiloAstro (89)	0	0	0	0	0	0	0	3	0	3	11	0	1	0	0	1	0	0	0	80	0	0	0	0
78	Uterus–AdenoCA (40)	2	0	0	5	0	10	0	0	0	0	0	0	0	2	2	0	0	0	0	0	78	0	0	0
74	Lung–AdenoCA (38)	3	0	0	0	0	0	0	3	5	0	0	0	3	0	3	0	0	3	0	5	0	74	0	3
73	Bone–Osteosarc (44)	0	0	0	2	0	2	0	2	2	5	0	0	7	0	0	0	0	0	0	5	0	0	73	2
61	Stomach–AdenoCA (70)	0	0	0	1	4	0	0	3	9	0	0	0	1	1	0	0	14	0	1	1	0	0	1	61

Fig. 2 Heatmap displaying the accuracy of the merged classifier using a held-out portion of the PCAWG data set for evaluation. Each row corresponds to the true tumour type; columns correspond to the class predictions emitted by the DNN. Cells are labelled with the percentage of tumours of a particular type that were classified by the DNN as a particular type. The recall and precision of each classifier are shown in the colour bars at the top and left sides of the matrix. All values represent the mean of 10 runs using selected data set partitions. Due to rounding of values, some rows add up to slightly more or less than 100%.

Results (2/4)

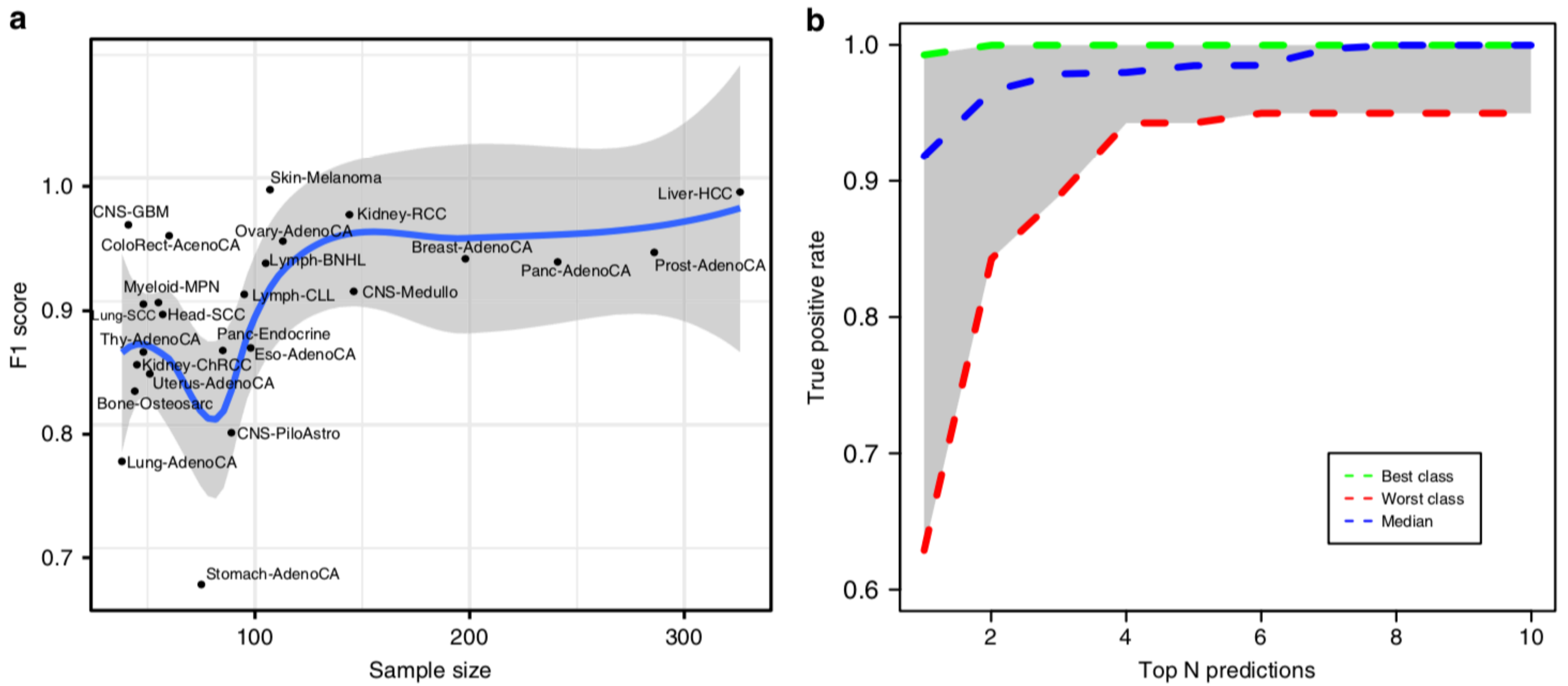
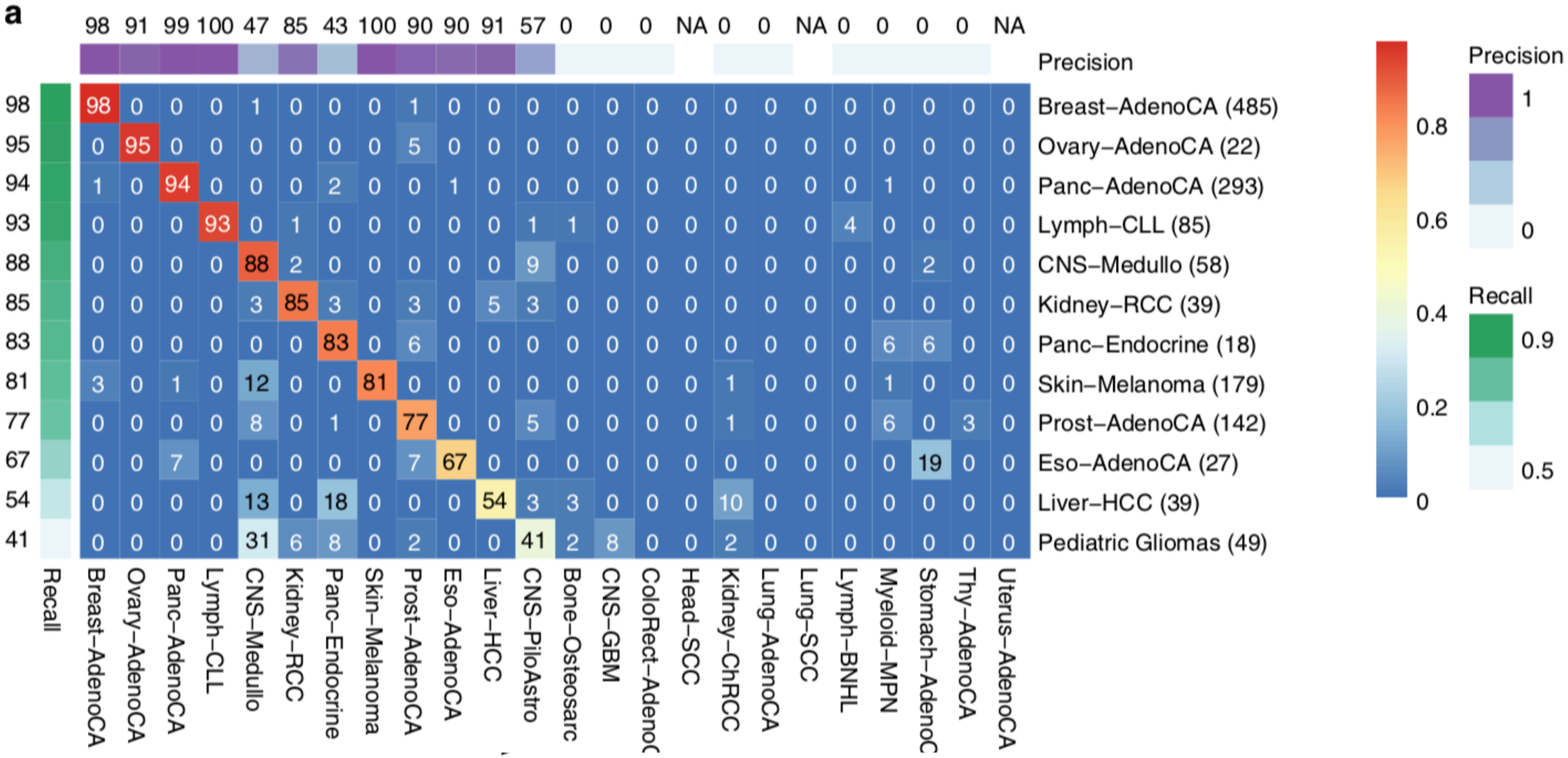


Fig. 3 Performance of the DNN on held-out PCAWG data. a The relationship between training set size and prediction accuracy of the DNN is shown for each tumour type. The blue line represents a regression line fit using LOESS regression, while the grey area represents a 95% confidence interval for the regression function. **b** Accuracy of the classifier when it is asked to identify the correct tumour type among its top N-ranked predictions. The blue dashed line is the median true-positive rate among all 24 tumour classes. The green and red dashed lines correspond to the true- positive rate for the best- and worst-performing tumour classes.

Results (3/4)

we applied the classifier trained on PCAWG sam- ples to an independent validation set of 1436 cancer whole genomes assembled from a series of published non-PCAWG projects. The validation set spans 14 distinct tumour types assembled from 21 publications or databases (Supplementary Data 4)

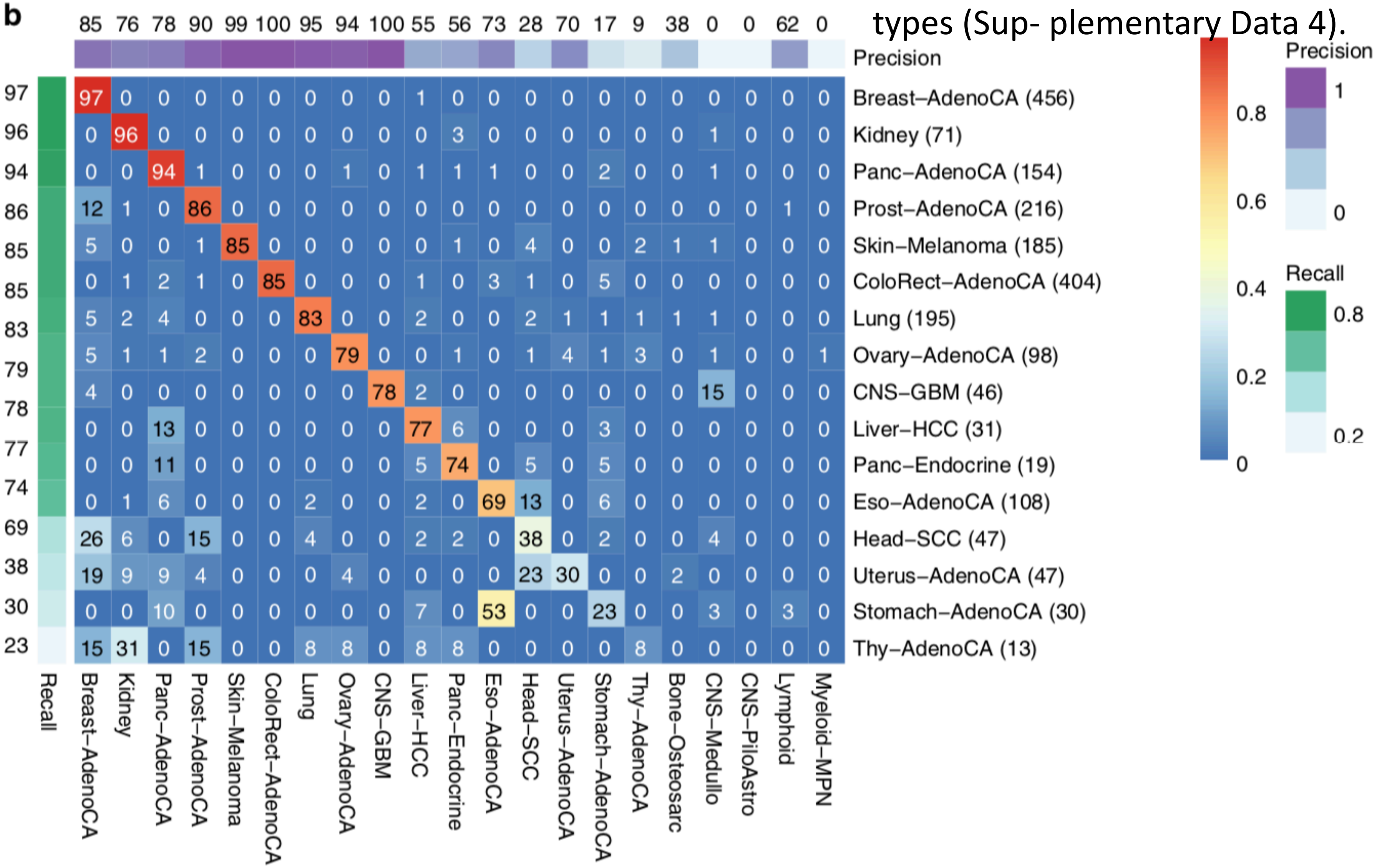


Recall	Breast–AdenoCA	Ovary–AdenoCA	Panc–AdenoCA	Lymph–CLL	CNS–Medullo	Kidney–RCC	Panc–Endocrine	Skin–Melanoma	Prost–AdenoCA	Eso–AdenoCA	Liver–HCC	CNS–PiloAstro	Bone–Osteosarc	CNS–GBM	ColoRect–AdenoC	Head–SCC	Kidney–ChRCC	Lung–AdenoCA	Lung–SCC	Lymph–BNHL	Myeloid–MPN	Stomach–AdenoC	Thy–AdenoCA	Uterus–AdenoCA	
	98	91	99	100	47	85	43	100	90	90	91	57	0	0	0	NA	0	0	NA	0	0	0	0	NA	Precision
98	98	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Breast–AdenoCA (485)
95	0	95	0	0	0	0	0	0	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	Ovary–AdenoCA (22)
94	1	0	94	0	0	0	2	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	Panc–AdenoCA (293)
93	0	0	0	93	0	1	0	0	0	0	0	1	1	0	0	0	0	0	0	4	0	0	0	0	Lymph–CLL (85)
88	0	0	0	0	88	2	0	0	0	0	0	9	0	0	0	0	0	0	0	0	0	2	0	0	CNS–Medullo (58)
85	0	0	0	0	3	85	3	0	3	0	5	3	0	0	0	0	0	0	0	0	0	0	0	0	Kidney–RCC (39)
83	0	0	0	0	0	0	83	0	6	0	0	0	0	0	0	0	0	0	0	0	6	6	0	0	Panc–Endocrine (18)
81	3	0	1	0	12	0	0	81	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	Skin–Melanoma (179)
77	0	0	0	0	8	0	1	0	77	0	0	5	0	0	0	0	1	0	0	0	6	0	3	0	Prost–AdenoCA (142)
67	0	0	7	0	0	0	0	0	7	67	0	0	0	0	0	0	0	0	0	0	0	19	0	0	Eso–AdenoCA (27)
54	0	0	0	0	13	0	18	0	0	0	54	3	3	0	0	0	10	0	0	0	0	0	0	0	Liver–HCC (39)
41	0	0	0	0	31	6	8	0	2	0	0	41	2	8	0	0	2	0	0	0	0	0	0	0	Pediatric Gliomas (49)

Fig. 4 Prediction accuracy for the DNN against two independent validation data sets. a Primary tumours. **b** Metastatic tumours. Each row corresponds to the true tumour type; columns correspond to the class predictions emitted by the DNN. Cells are labelled with the percentage of tumours of a particular type that were classified by the DNN as a particular type. The recall and precision of each classifier are shown in the colour bars at the top and left sides of the matrix. Due to rounding of values, some rows add up to slightly more or less than 100%.

Results (4/4)

that combined a published series of 92 metastatic Panc-AdenoCA25 with an unpublished set of 2,028 metastatic tumours from known primaries across 16 tumour types recently sequenced by the Hartwig Medical Foundation (HMF), resulting in a combined set of 2120 samples across 16 tumour types (Sup- plementary Data 4).



	Recall	Breast–AdenoCA	Kidney	Panc–AdenoCA	Prost–AdenoCA	Skin–Melanoma	ColoRect–AdenoCA	Lung	Ovary–AdenoCA	CNS–GBM	Liver–HCC	Panc–Endocrine	Eso–AdenoCA	Head–SCC	Uterus–AdenoCA	Stomach–AdenoCA	Thy–AdenoCA	Bone–Osteosarc	CNS–Medullo	CNS–PiloAstro	Lymphoid	Myeloid–MPN
Precision		85	76	78	90	99	100	95	94	100	55	56	73	28	70	17	9	38	0	0	62	0
Breast–AdenoCA (456)	97	97	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0
Kidney (71)	96	0	96	0	0	0	0	0	0	0	0	3	0	0	0	0	0	0	1	0	0	0
Panc–AdenoCA (154)	94	0	0	94	1	0	0	0	1	0	1	1	1	0	0	2	0	0	1	0	0	0
Prost–AdenoCA (216)	86	12	1	0	86	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0
Skin–Melanoma (185)	85	5	0	0	1	85	0	0	0	0	0	1	0	4	0	0	2	1	1	0	0	0
ColoRect–AdenoCA (404)	85	0	1	2	1	0	85	0	0	0	1	0	3	1	0	5	0	0	0	0	0	0
Lung (195)	83	5	2	4	0	0	0	83	0	0	2	0	0	2	1	1	1	1	1	0	0	0
Ovary–AdenoCA (98)	79	5	1	1	2	0	0	0	79	0	0	1	0	1	4	1	3	0	1	0	0	1
CNS–GBM (46)	78	4	0	0	0	0	0	0	0	78	2	0	0	0	0	0	0	0	15	0	0	0
Liver–HCC (31)	77	0	0	13	0	0	0	0	0	0	77	6	0	0	0	3	0	0	0	0	0	0
Panc–Endocrine (19)	74	0	0	11	0	0	0	0	0	0	5	74	0	5	0	5	0	0	0	0	0	0
Eso–AdenoCA (108)	69	0	1	6	0	0	0	2	0	0	2	0	69	13	0	6	0	0	0	0	0	0
Head–SCC (47)	38	26	6	0	15	0	0	4	0	0	2	2	0	38	0	2	0	0	4	0	0	0
Uterus–AdenoCA (47)	30	19	9	9	4	0	0	0	4	0	0	0	0	23	30	0	0	2	0	0	0	0
Stomach–AdenoCA (30)	23	0	0	10	0	0	0	0	0	0	7	0	53	0	0	23	0	0	3	0	3	0
Thy–AdenoCA (13)		15	31	0	15	0	0	8	8	0	8	8	0	0	0	0	8	0	0	0	0	0