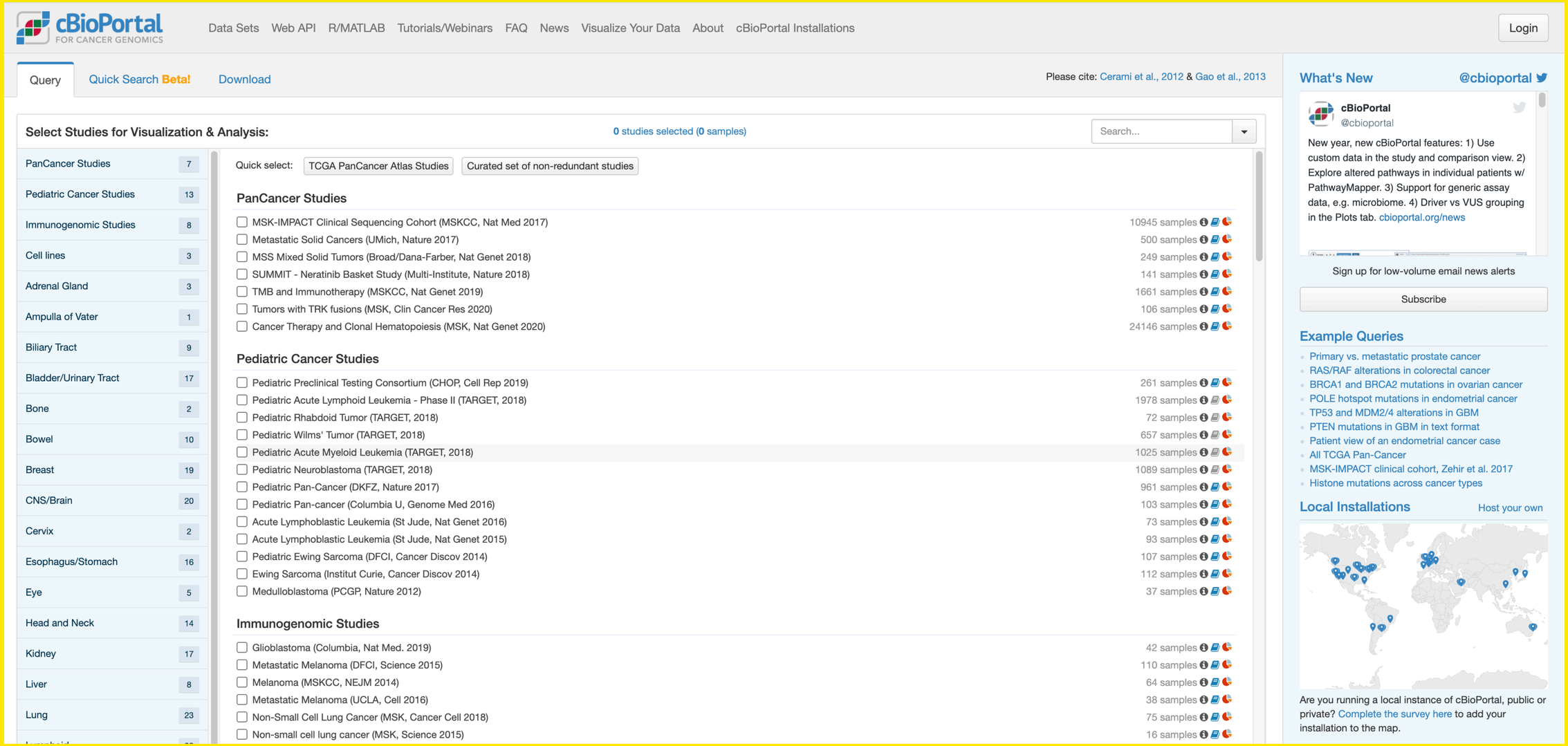
Prediction of Cancer Type

BASED ON CLINICAL AND GENOMIC FEATURES

HARMEET GILL

GA CAPSTONE

CBIOPORTAL FOR CANCER GENOMICS



CBIOPORTAL FOR CANCER GENOMICS

An open-source resource for cancer genomic datasets

The Cancer Genome Atlas (TCGA) consortium

Data collection

AVAILABLE DATA FROM CBIOPORTAL

303

Studies

118K+

Samples

869

Cancers

FILTERED DATA FOR AT LEAST 2000 SAMPLES

129

Features

85K

Samples

154

Cancers

EDA

Overview

Select subsets

2000 samples

Merge datasets

ON sample ID | unique features

Drop

Redundant features | samples with no cancer type | potential targets

---

Clean values

.unique | .dtypes | .ISALPHA

Binning

Range of values | Fill missing values

EDA

Redundant features

- Oncotree code
- Study ID
- Other sample ID
- Specimen preservation type
- Pathology report file number
- Patient ID
- Vial number
- FFPE
- OCT embedded
- Analysis cohort
- Institute

EDA

Cleaning example

jupyter 2_capstone_notebook_cleaning_EDA Last Checkpoint: 2 hours ago (unsaved changes)

File Edit View Insert Cell Kernel Navigate Widgets Help Trusted Python 3

Run Code Snippets

7.0.1 cleaning ploidy

In [52]:
```
pd.DataFrame(final_df['PLOIDY'].unique())
```

Out[52]:

	0
0	NaN
1	3.078991415
2	1.871227842
3	3.521216365
4	2.407595488
5	2.53804551
6	3.272776561
7	1.952094725
8	1.990020258
9	3.246111178
10	2.197296274

In [ ]:
```
final_df['PLOIDY'].replace('3n+, Hypertriploid (70-80)', 'unknown', inplace=True)
final_df['PLOIDY'].replace('4n-, Hypotetraploid (81-91)', 'unknown', inplace=True)
final_df['PLOIDY'].replace('5n+/-, Near-pentaploid 115+/- (104-126)', 'unknown', inplace=Tr
final_df['PLOIDY'].replace('4n, Tetraploid (92)', 'unknown', inplace=True)
final_df['PLOIDY'].replace('3n-, Hypotriploid (58-68)', 'unknown', inplace=True)

final_df['PLOIDY'].replace('2n+, Hyperdiploid (47-57)', 'unknown', inplace=True)
final_df['PLOIDY'].replace('4n+, Hypertetraploid (93-103)', 'unknown', inplace=True)
final_df['PLOIDY'].replace('2n-, Hypodiploid (35-45)', 'unknown', inplace=True)
final_df['PLOIDY'].replace('2n+/-, Near-diploid 46+/- (35-57)', 'unknown', inplace=True)

final_df['PLOIDY'].replace('3n+/-, Near-triploid 69+/- (58-80)', 'unknown', inplace=True)
final_df['PLOIDY'].replace('3n, Triploid (69)', 'unknown', inplace=True)
final_df['PLOIDY'].replace('4n+/-, Near-tetraploid 92+/- (81-103)', 'unknown', inplace=True
```

21

Continuous features

'FRACTION_GENOME_ALTERED',
'MUTATION_COUNT',
'DNA_INPUT',
'SAMPLE_COVERAGE',
'MSI_SENSOR_SCORE',
'ANEUPLOIDY_SCORE',
'MSI_SCORE_MANTIS',
'SAMPLE_INITIAL_WEIGHT',
'OVERALL_SURVIVAL_MONTHS',
'NUMBER_OF_SAMPLES_PER_PATIENT',
'Age',
'Number of Samples Per Patient',
'Time from Diagnosis',
'Time to Blood Draw from Treatment'

101

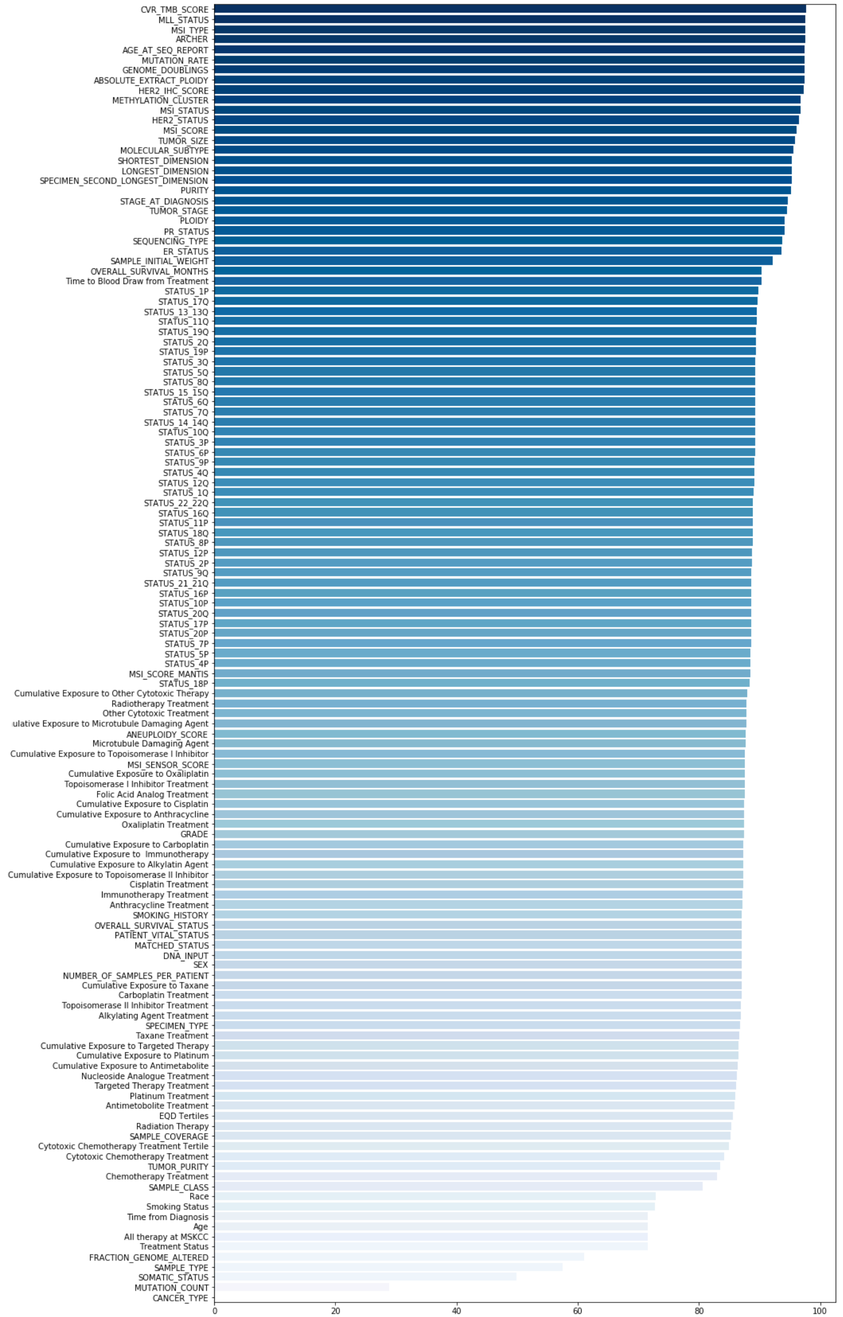
Categorical features

'SAMPLE_TYPE',
'SOMATIC_STATUS',
'METASTATIC_SITE',
'SAMPLE_CLASS',
'SEQUENCING_TYPE',
'GRADE',
'ER_STATUS',
'TUMOR_PURITY',
'SPECIMEN_TYPE',
'MATCHED_STATUS',
'OVERALL_SURVIVAL_STATUS',
'SEX',
'SMOKING_HISTORY',
'PATIENT_VITAL_STATUS',
'Alkylating Agent Treatment',
AND 85 MORE…

7

Targets

‘PRIMARY_SITE',
‘PRIMARY_TUMOR_SITE',
‘TUMOR_TYPE',
‘TUMOR_TISSUE_SITE',
‘CANCER_TYPE_DETAILED',
‘METASTATIC_SITE',
‘TISSUE_SOURCE_SITE',



78-93% NULLS

EDA: missing values

Of the 303 studies, there is little overlap between the reported features, resulting in missing values.

However, the minimum number of samples for any given feature is 2000.

EDA: Binning

Missing values in features with continuous data could not be removed or filled without distorting the data.

Therefore, the data was put into bins.

jupyter 2_capstone_notebook_cleaning_EDA Last Checkpoint: 3 hours ago (autosaved)

Logout

File Edit View Insert Cell Kernel Navigate Widgets Help

Trusted Python 3

Run Code Snippets

8.3 binning variables with 'unknown' values

In [103]:
```
print(final_df['TUMOR_SIZE'].unique())
```

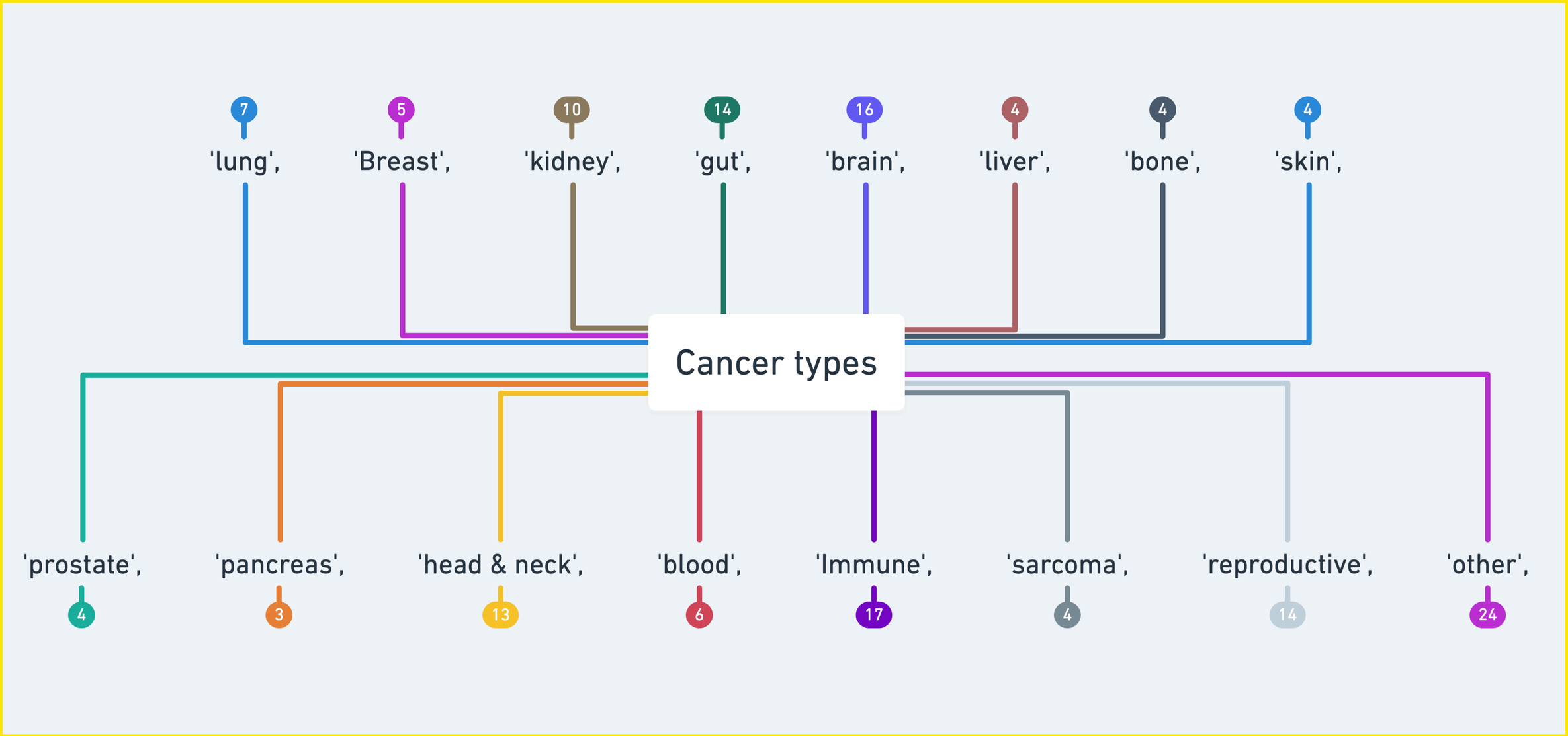
```
[nan '7.5' '5' '6' '4.5' '2' '4' '3.5' '5.5' '3' '2.5' '1.7' '1' '7' '1.5'
 '4.6' '0.8' '12' '2.2' '0.2' '3.2' '3.8' '6.5' '3.6' '3.7' '15' '1.1'
 '1.4' '2.7' '25' '24' '20' '35' '23' '40' '22' '17' '45' '30' '60' '38'
 '26' '70' '18' '65' '50' '80' '100' '130' '120' '32' '180' '140' '63'
 '650' '90' '67' '95' '28' '48' '190' '44' '220' '200' '75' '160' '16'
 '150' '55' '85' '210' '105' '36' '11' '110' '58' '33' '170' '10' '142'
 '34' '57' '8' '135' '52' '47' '43' '13' '19' '27' '71' '14' '21' '37'
 '42' '115' '39' '29' '72' '0.7' '4.2' '2.8' '1.2' '9' 'unknown' '1.8'
 '3.4' '2.3' '2.6' '2.4' '5.3' '5.4' '4.8' '5.7' '7.2' '5.8' '5.2' '2.9'
 '3.9' '4.1' 63.0 30.0 24.0 60.0 18.0 22.0 45.0 80.0 87.0 20.0 35.0 43.0
 25.0 21.0 40.0 1.5 10.0 15.0 31.0 65.0 29.0 34.0 16.0 28.0 19.0 36.0 33.0
 23.0 17.0 12.0 50.0 13.0 14.0 55.0 39.0 70.0 27.0 150.0 26.0 9.0 38.0 2.0
 52.0 44.0 48.0 3.0 5.0 46.0 11.0 53.0 47.0 32.0 67.0 42.0 180.0 57.0 4.0
 100.0 37.0 90.0 8.0 160.0 84.0 130.0 5.5 62.0 1.0 49.0 99.0 68.0 7.0 41.0
 6.0 75.0 51.0 120.0 61.0 79.0 71.0 22.5 17.9 14.5 12.8 18.5 15.5 21.5
 16.9 24.4 12.5 40.3 11.8 32.6 17.2 13.8 15.7 182.0 85.0 18.3 21.6 28.5
 16.2 2.3 15.2 31.1 14.3 12.6 25.1 17.6 2.12 21.3 22.32 17.7 15.47 24.15
 20.5 69.0 88.0 13.6 66.0 20.8 27.3 21.1 22.8 25.2 26.7 8.9 19.1 11.4 14.9
 20.33 17.06 18.95 31.7 21.87 24.6 24.05 1.7 17.3 21.4 19.5 58.0 76.0
 105.0 110.0 2.7 0.7 2.5 1.8 1.4 3.5 1.9 3.2 1.1 4.5 1.3 1.2 2.8 2.2 8.4
 4.1 2.1 0.0 3.8 11.5 6.5 7.9 2.6 9.5 24.5 39.5 8.5 13.5 13.2 18.2 10.6
 9.2 6.4 6.6 3.7 6.3 4.2 38.3 10.8 4.4 7.2 7.1 9.7 8.8 12.9 5.6 10.9 12.4
 27.5 15.1 7.5 5.4 8.1 10.5 13.7 4.9 18.8 3.9 5.1 36.5 7.6 10.2 29.5 17.5
 11.1 15.3 1.6 0.6 2.4 0.5 '53' '74' '86' '93' '92' '126' '98' '41' '51'
 '76' '82' '46' '68' '186' '91' '79' '81' '112' '107' '83' '31' '121' '78']
```

In [104]:
```
def create_bins_with_unknowns(df, col, bins):
    df[col].replace('unknown', np.nan, inplace=True)
    df[col].replace('Unknown', np.nan, inplace=True)
    df[col]=df[col].astype(float)
    df[col]=pd.cut(df[col], bins=bins, include_lowest=True)
    df[col]=df[col].cat.add_categories('unknown')
    df[col]=df[col].fillna('unknown')
    return df[col].value_counts()
```

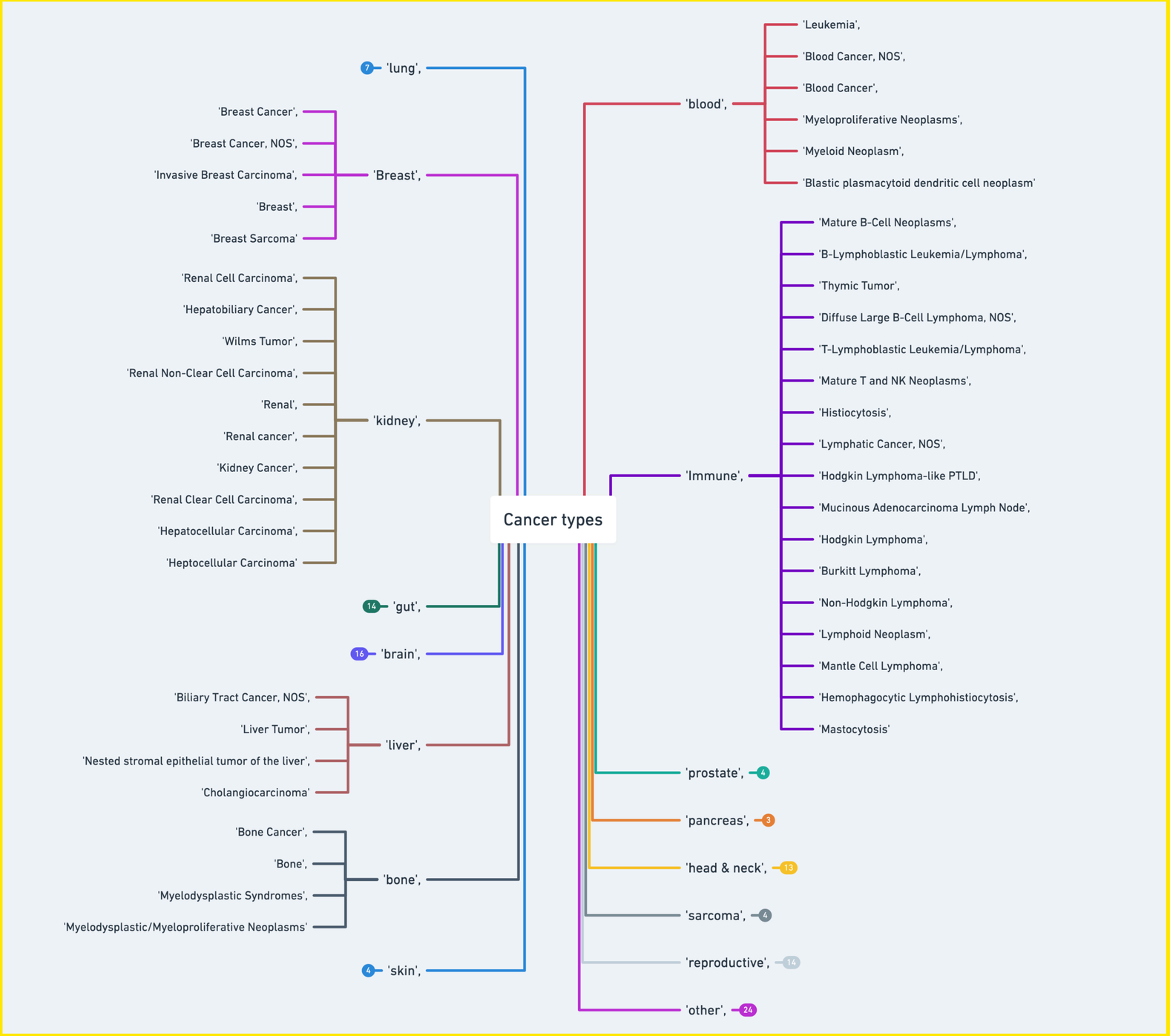
In [105]:
```
create_bins_with_unknowns(final_df, 'TUMOR_SIZE', bins = [-0.6, 10, 20, 30, 40, 50, 60, 650
```

Out[105]:
```
unknown           81184
(10.0, 20.0]       1053
(20.0, 30.0]        918
(-0.601, 10.0]      499
(30.0, 40.0]        374
(60.0, 650.0]       264
(40.0, 50.0]        210
(50.0, 60.0]        107
Name: TUMOR_SIZE, dtype: int64
```

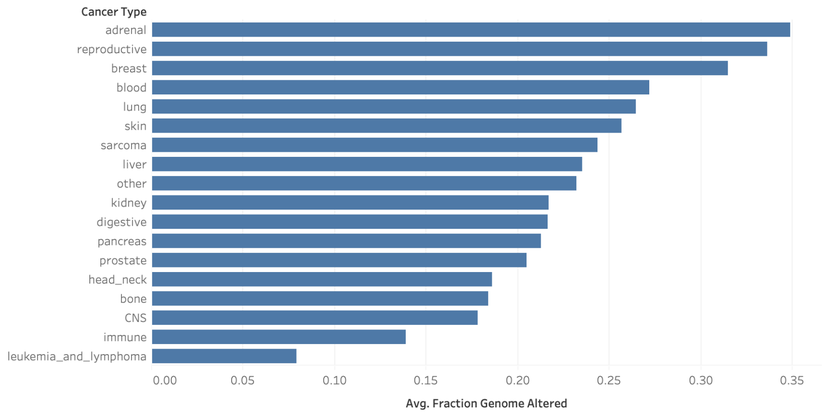
Categorised cancer types



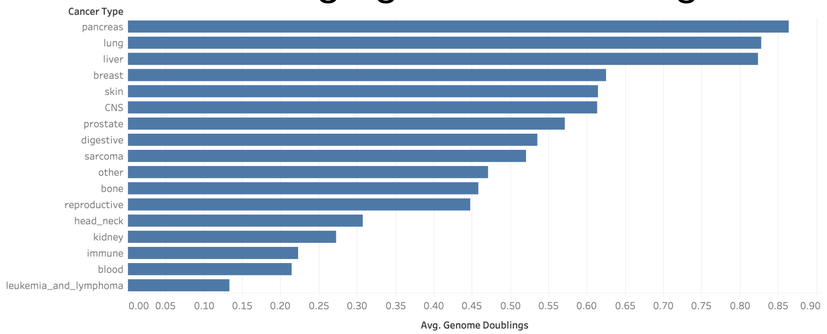
Categorised cancer types



Average fraction genome altered



Average genome doublings



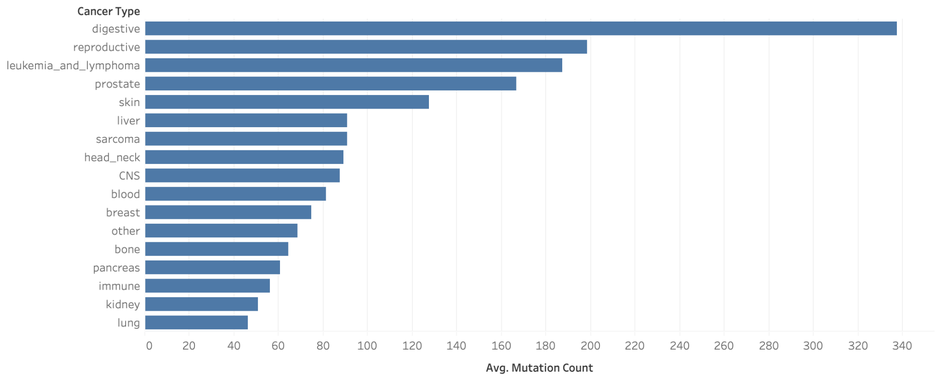
EDA: Genomic insights

Lymphomas, **leukaemias**, cancers related to the **cardiovascular** and **immune** systems contain one of the **lowest incidence** of genomic changes (i.e. copy number alterations).

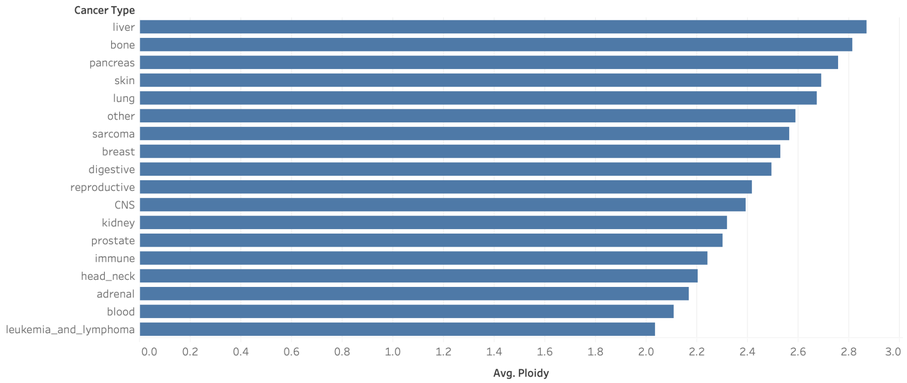
This indicates relatively **less genomic instability** and environment of **tumour promotion** as compared to e.g. breast cancer.

EDA: Genomic insights

Average mutation count



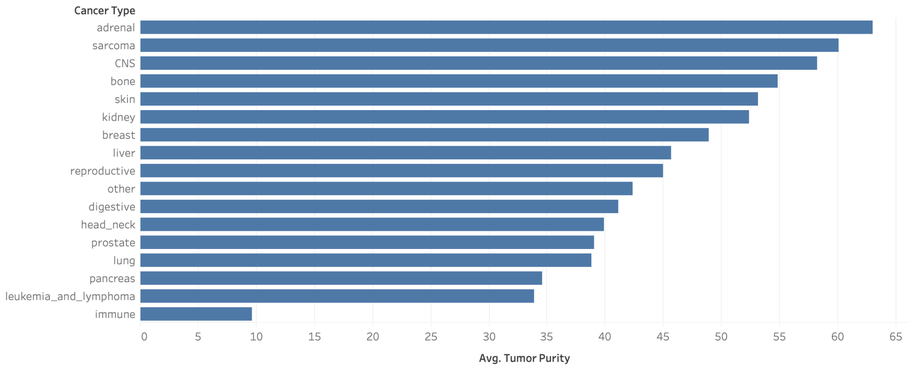
Average ploidy



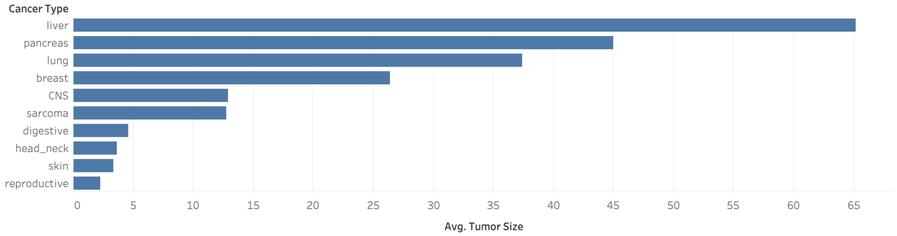
However, **mutation counts** in **lymphomas** and **leukaemias** are comparatively **high.**

In contrast, cancers of the **digestive system** display the **opposite pattern.**

Average tumour purity



Average tumour size (cm)



EDA: Tumour insights

Tumour purity (the **proportion of cancerous cells**) - cancers of the **immune system** displayed the **lowest levels**, while the remaining cancers could be split roughly into 2 groups (>45% and <45% tumour purity).

Tumour size can only be reported for **solid tumours** and interestingly, the average tumour size (cm) was **highest** for tumours of the **liver** in **striking contrast** to tumours of the **digestive**, **reproductive** system, **head/neck** and **skin.**

Average age by race

Cancer Type	Race: Asian	Black	other	White
adrenal	40.81	47.19	60.59	53.53
bone	29.78	34.57	26.95	39.56
breast	51.84	55.35	53.01	57.29
CNS	46.64	43.76	45.96	51.98
digestive	56.55	57.79	54.76	60.27
head_neck	56.25	61.53	54.95	61.08
immune	49.61	52.50	67.41	60.60
kidney	53.33	46.52	50.91	56.78
liver	62.37	66.15	58.57	54.89
lung	63.10	62.36	62.32	67.21
other	46.49	47.27	44.00	54.05
pancreas	61.23	61.16	57.73	65.14
prostate	66.07	64.66	62.47	67.40
reproductive	57.71	62.00	56.60	62.09
sarcoma	48.05	45.20	38.77	52.74
skin	57.48	59.19	62.53	63.92

Avg. Age

26.95 67.41

EDA: Clinical insights

Across different races, there are **similarities** in the average age of cancer patients (e.g. **prostate and lung**), while **large differences** lie in **adrenal**, **breast** and **liver** cancers.

Modelling

Model 1	Logistic regression
Model 2	Decision tree classifier
Model 3	Random forest

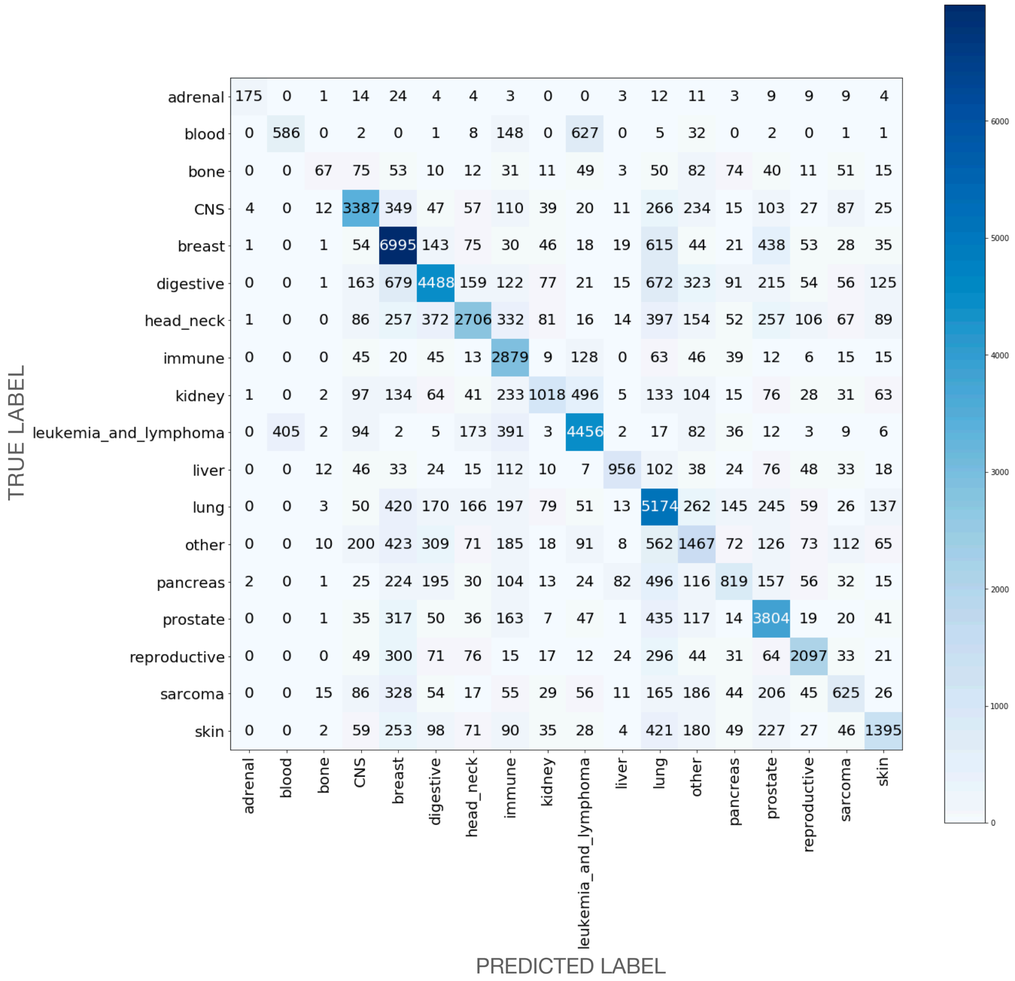
Logistic regression

MODEL 1

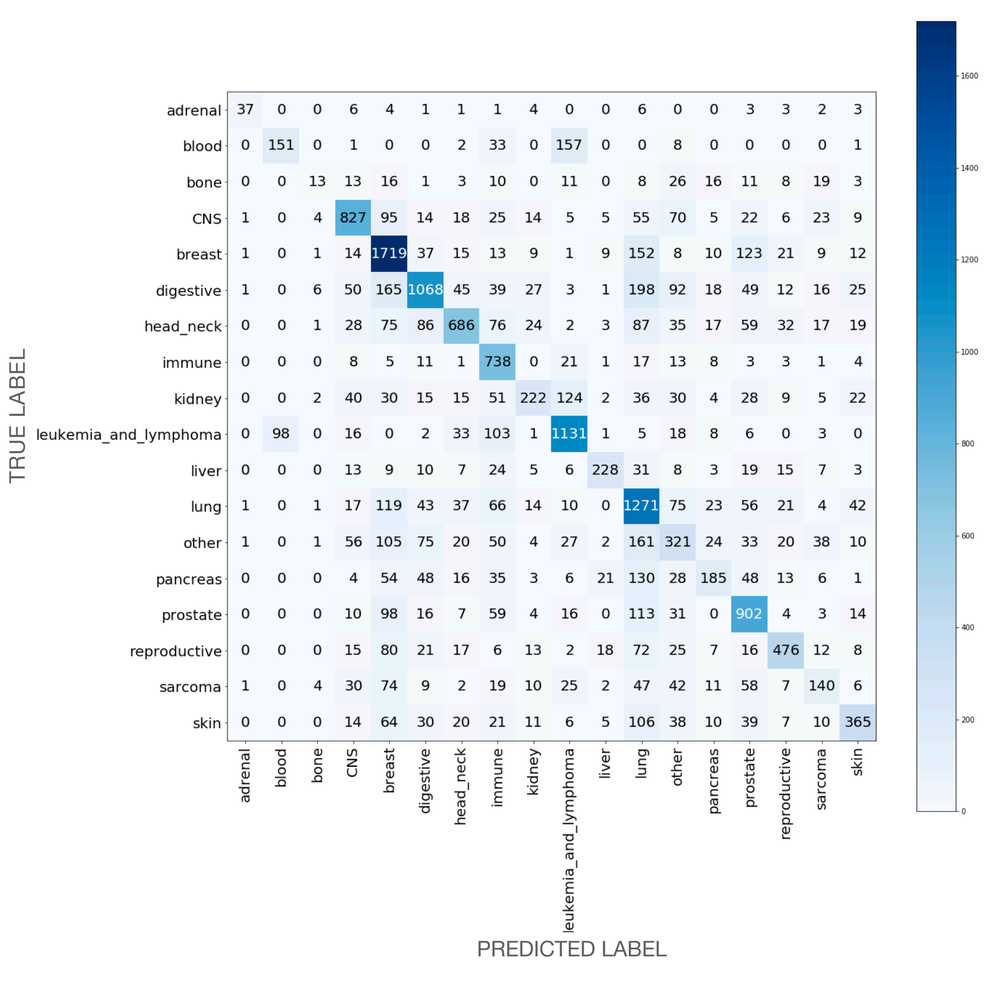
0.13	Baseline
0.62	Best score
0.62	Test score
0.64	Train score

Model 1: Logistic Regression - confusion matrices

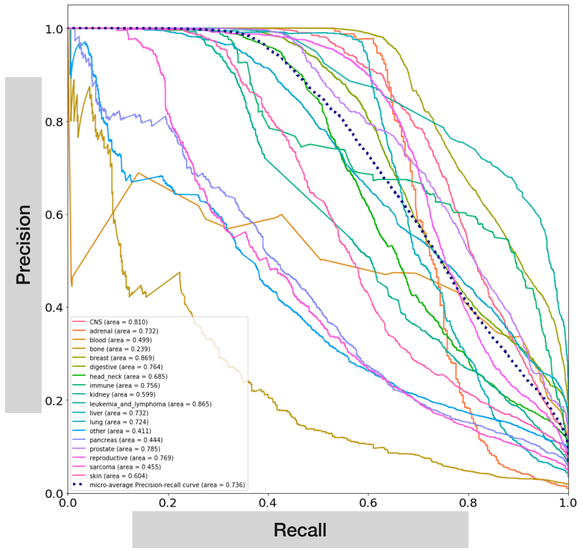
TRAINING SET

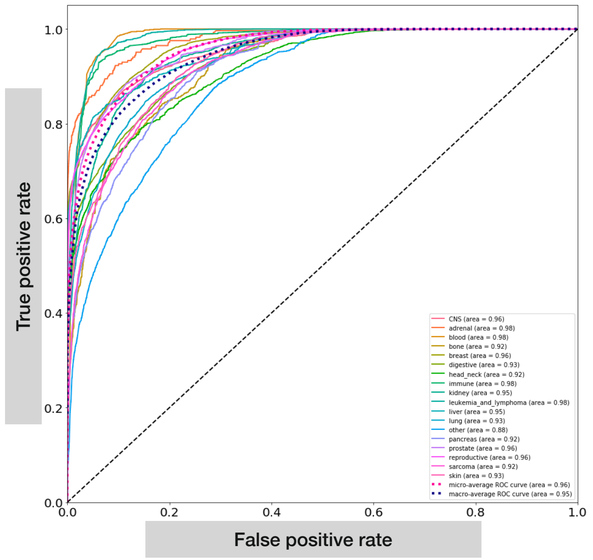


TEST SET



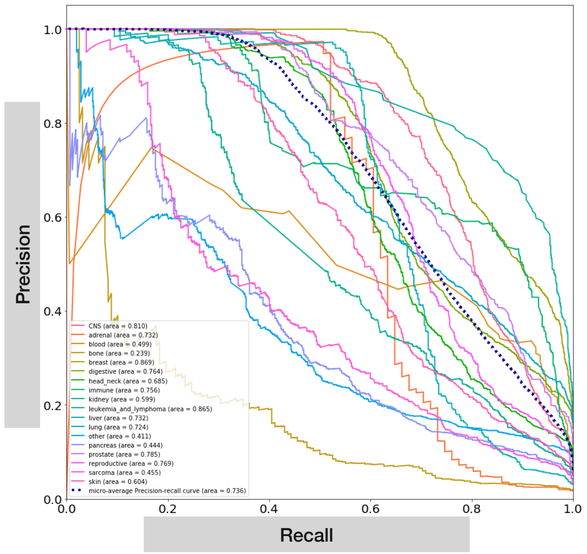
Model 1: Logistic Regression - classification report for training set

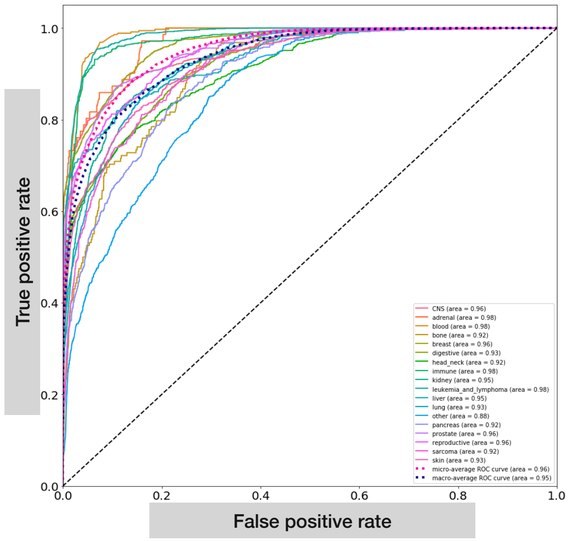




	precision	recall	f1-score	support
CNS	0.74	0.71	0.72	4793
adrenal	0.95	0.61	0.75	285
blood	0.59	0.41	0.49	1413
bone	0.52	0.11	0.18	634
breast	0.65	0.81	0.72	8616
digestive	0.73	0.62	0.67	7261
head_neck	0.73	0.54	0.62	4987
immune	0.55	0.86	0.67	3335
kidney	0.68	0.40	0.50	2541
leukemia_and_lymphoma	0.72	0.78	0.75	5698
liver	0.82	0.62	0.70	1554
lung	0.52	0.72	0.61	7197
other	0.42	0.39	0.40	3792
pancreas	0.53	0.34	0.42	2391
prostate	0.63	0.74	0.68	5107
reproductive	0.77	0.67	0.71	3150
sarcoma	0.49	0.32	0.39	1948
skin	0.67	0.47	0.55	2985
accuracy			0.64	67687
macro avg	0.65	0.56	0.59	67687
weighted avg	0.64	0.64	0.63	67687

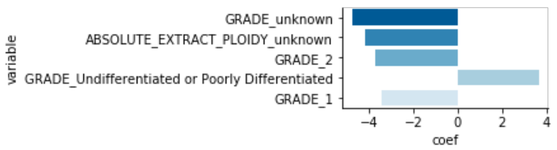
Model 1: Logistic Regression - classification report for test set



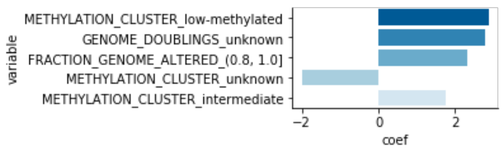


	precision	recall	f1-score	support
CNS	0.71	0.69	0.70	1198
adrenal	0.86	0.52	0.65	71
blood	0.61	0.43	0.50	353
bone	0.39	0.08	0.14	158
breast	0.63	0.80	0.71	2154
digestive	0.72	0.59	0.65	1815
head_neck	0.73	0.55	0.63	1247
immune	0.54	0.88	0.67	834
kidney	0.61	0.35	0.44	635
leukemia_and_lymphoma	0.73	0.79	0.76	1425
liver	0.77	0.59	0.66	388
lung	0.51	0.71	0.59	1800
other	0.37	0.34	0.35	948
pancreas	0.53	0.31	0.39	598
prostate	0.61	0.71	0.66	1277
reproductive	0.72	0.60	0.66	788
sarcoma	0.44	0.29	0.35	487
skin	0.67	0.49	0.56	746
accuracy			0.62	16922
macro avg	0.62	0.54	0.56	16922
weighted avg	0.63	0.62	0.61	16922

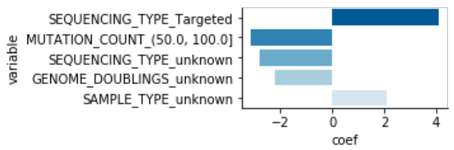
CNS



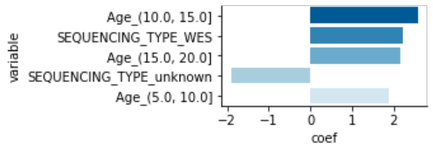
Adrenal



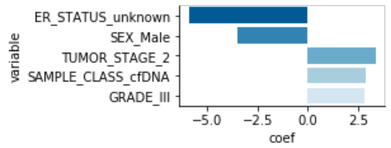
Blood



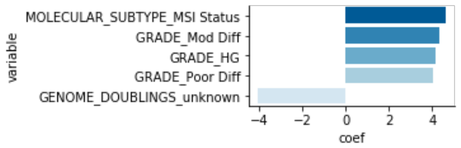
Bone



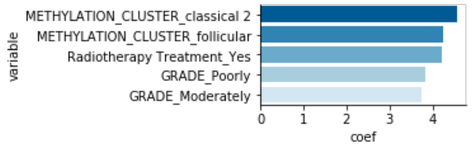
Breast



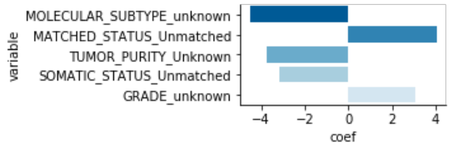
Digestive



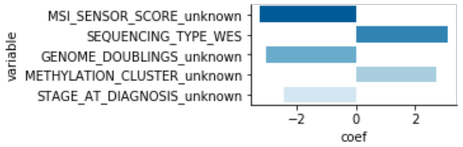
Head/neck



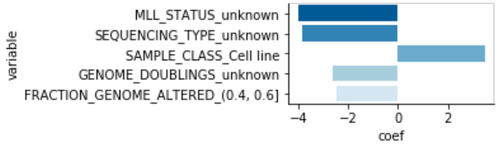
Immune



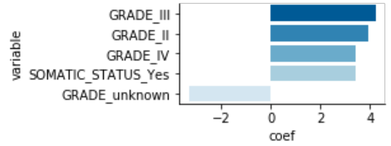
Kidney



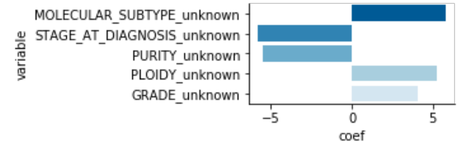
Leukaemia/lymphoma



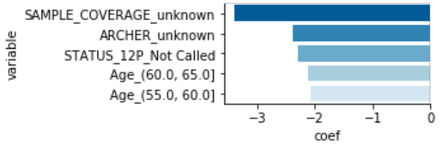
Liver



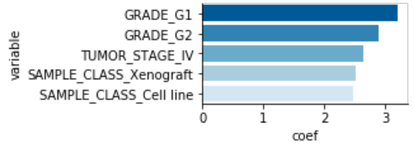
Lung



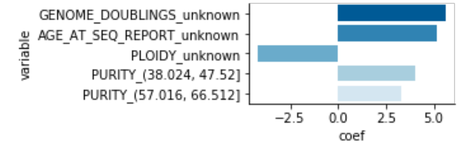
Other



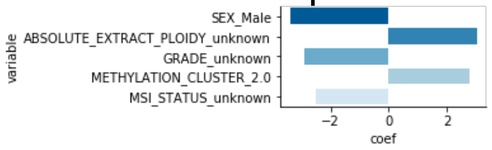
Pancreas



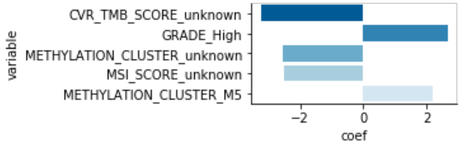
Prostate



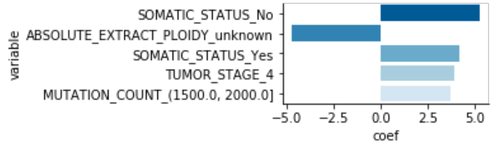
Reproductive



Sarcoma



Skin



Top 5 coefficients

Supporting earlier EDA, **leukaemias and lymphomas** are **negatively** correlated with **fraction genome altered**, while again, agreeing with EDA, **adrenal** cancers are **positively correlated** fraction genome altered.

Decision tree classifier

MODEL 2

0.62 Best score

0.63 Test score

0.68 Train score

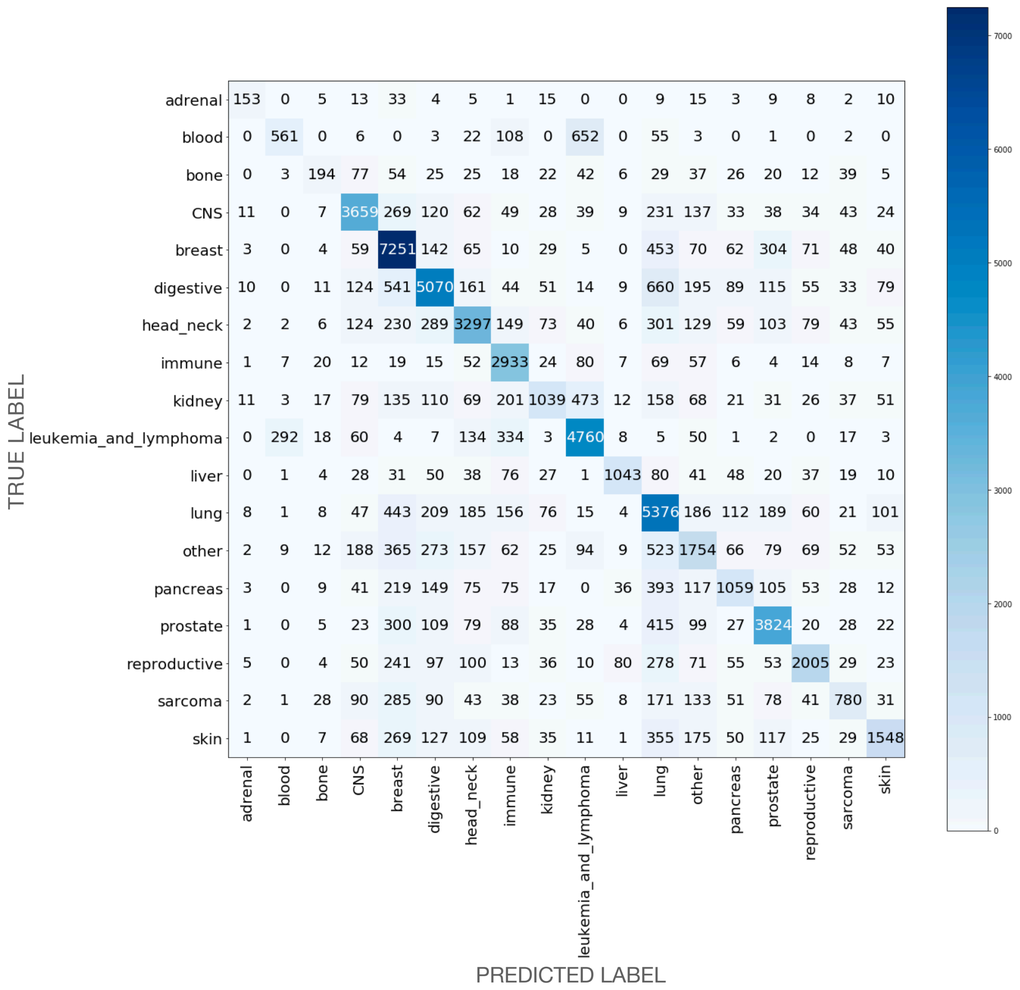
LOGISTIC REGRESSION

MODEL 1 SCORE

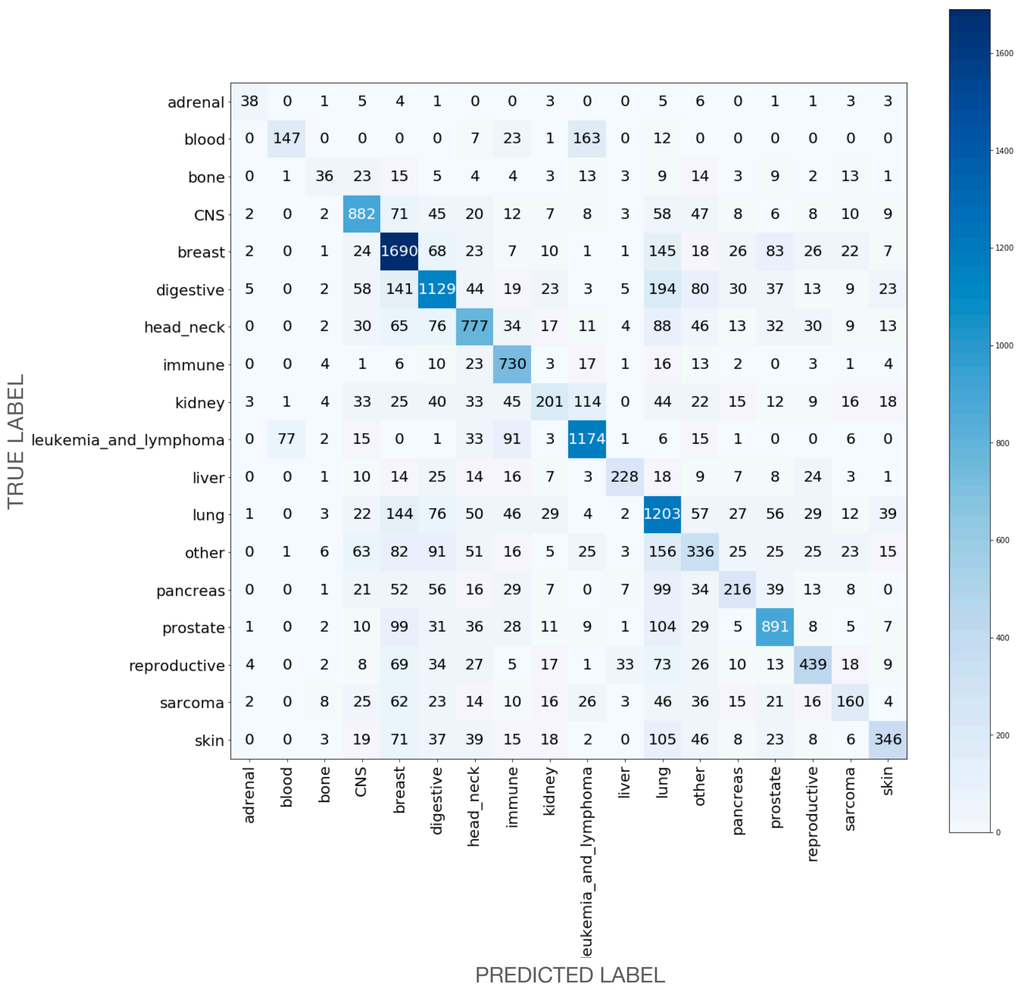
0.62	Best score
0.62	Test score
0.64	Train score

Model 2: Decision tree classifier - confusion matrices

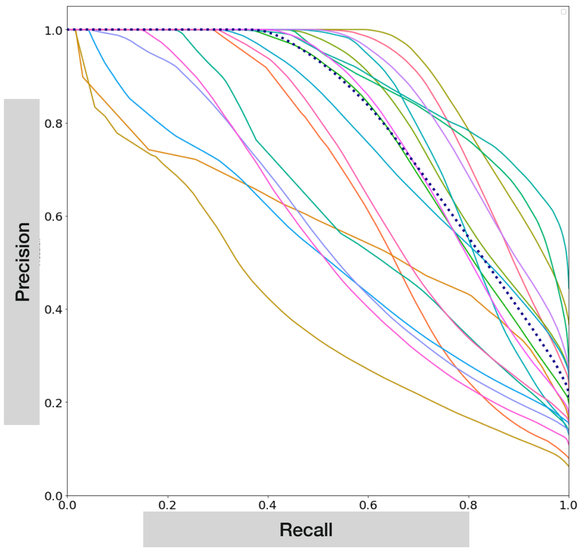
TRAINING SET

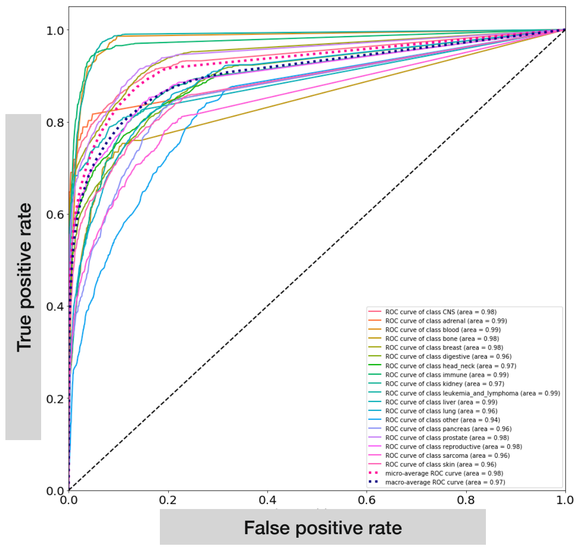


TEST SET



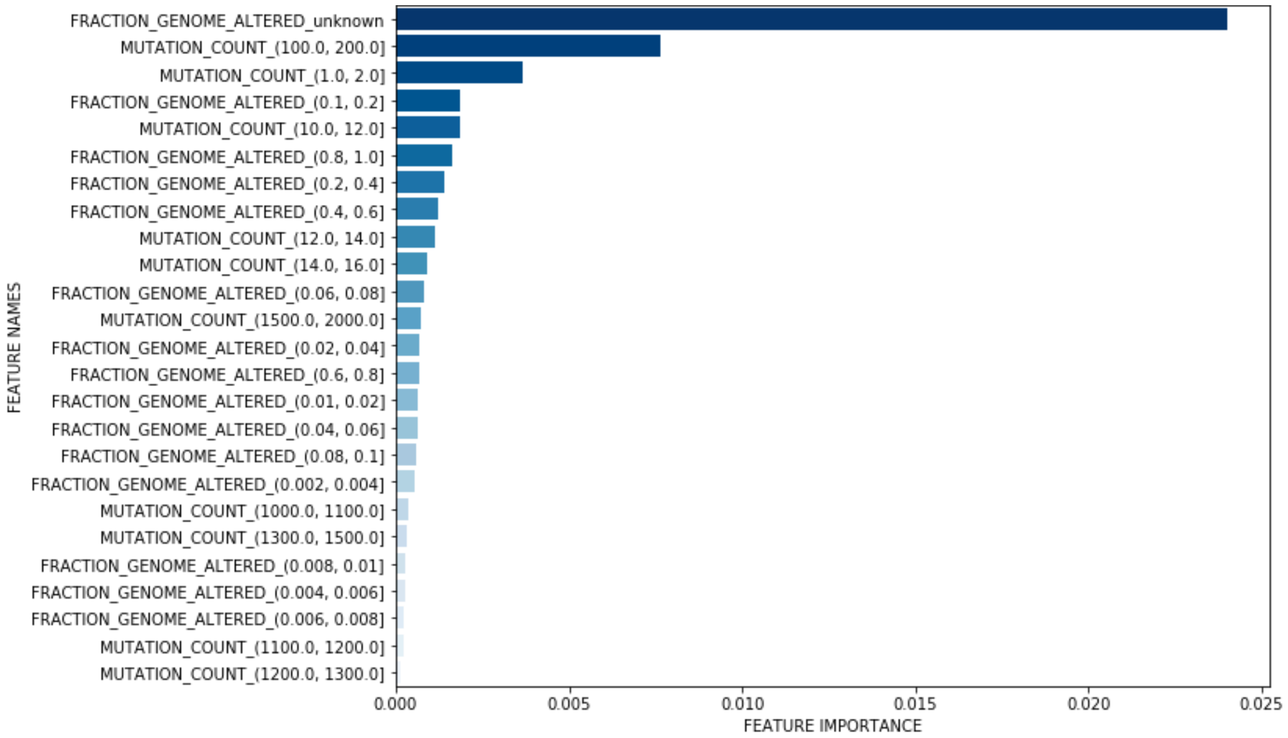
Model 2: Decision tree classifier - classification report for test set





	precision	recall	f1-score	support
CNS	0.71	0.74	0.72	1198
adrenal	0.66	0.54	0.59	71
blood	0.65	0.42	0.51	353
bone	0.45	0.23	0.30	158
breast	0.65	0.78	0.71	2154
digestive	0.65	0.62	0.63	1815
head_neck	0.64	0.62	0.63	1247
immune	0.65	0.88	0.74	834
kidney	0.53	0.32	0.40	635
leukemia_and_lymphoma	0.75	0.82	0.78	1425
liver	0.77	0.59	0.67	388
lung	0.51	0.67	0.58	1800
other	0.40	0.35	0.38	948
pancreas	0.53	0.36	0.43	598
prostate	0.71	0.70	0.70	1277
reproductive	0.67	0.56	0.61	788
sarcoma	0.49	0.33	0.39	487
skin	0.69	0.46	0.56	746
accuracy			0.63	16922
macro avg	0.62	0.55	0.57	16922
weighted avg	0.63	0.63	0.62	16922

Model 2: Decision tree classifier - top 25 features



Mutation count and **fraction genome altered** appear to be the most important features for predicting cancer types with decision tree classifier model.

RANDOM FOREST

MODEL 3

0.62 Best score

0.63 Test score

0.74 Train score

LOGISTIC REGRESSION

MODEL 1 SCORE

0.62	Best score
0.62	Test score
0.64	Train score

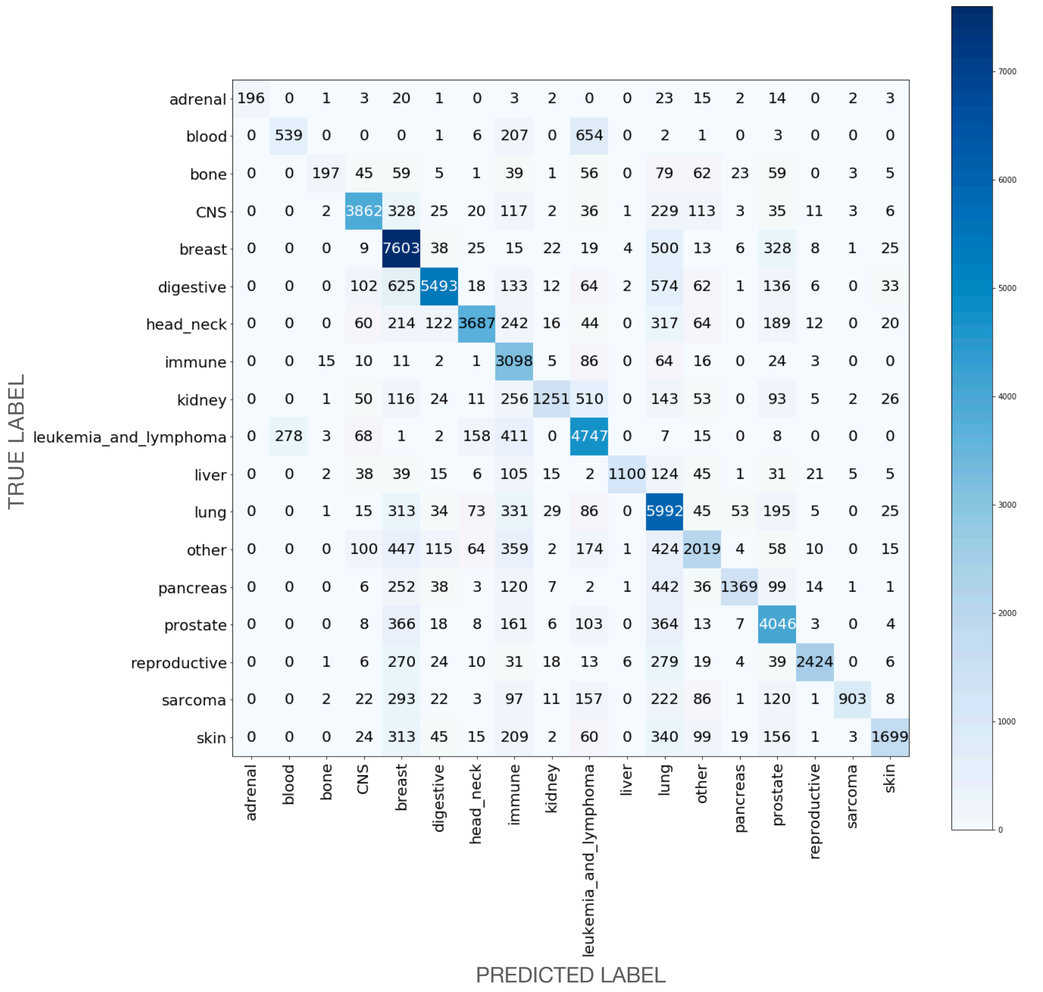
DECISION TREE CLASSIFIER

MODEL 1 SCORE

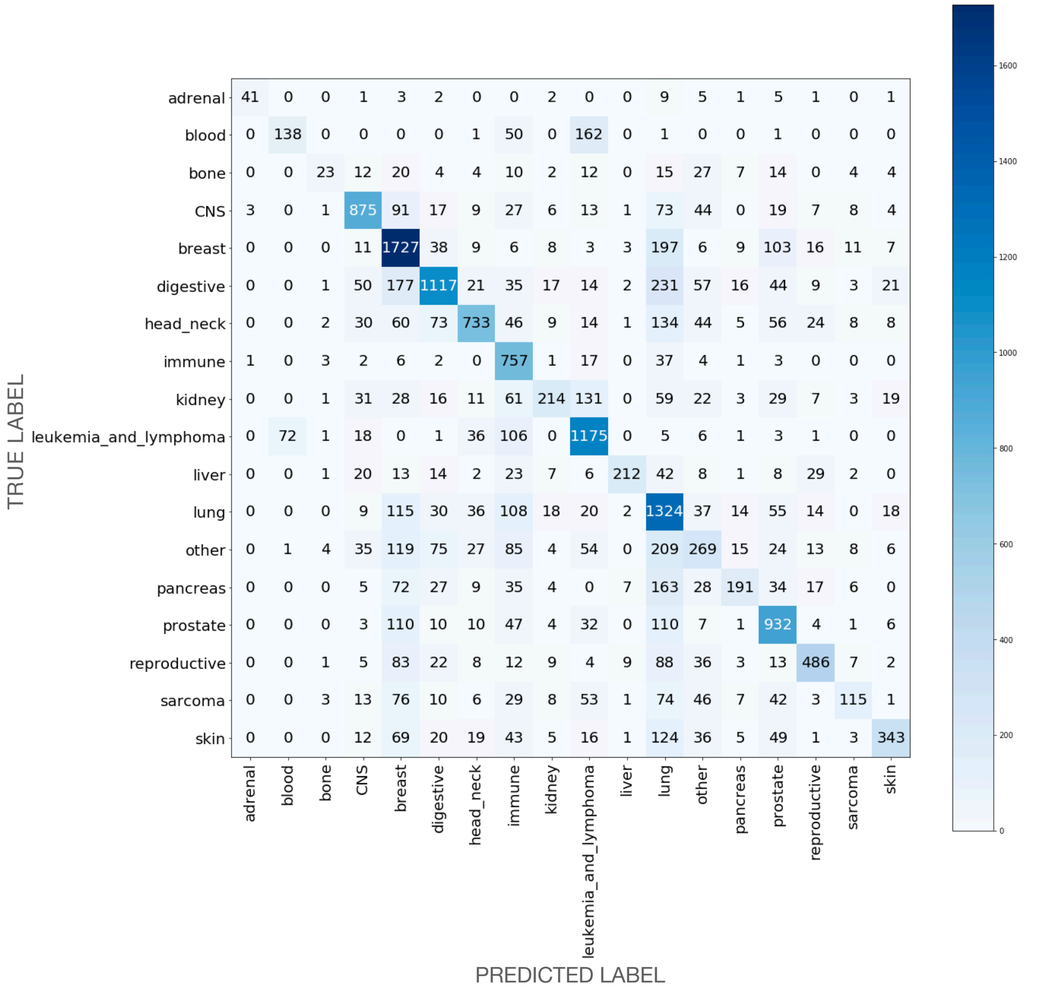
0.62	Best score
0.63	Test score
0.68	Train score

Model 2: Random forest - confusion matrices

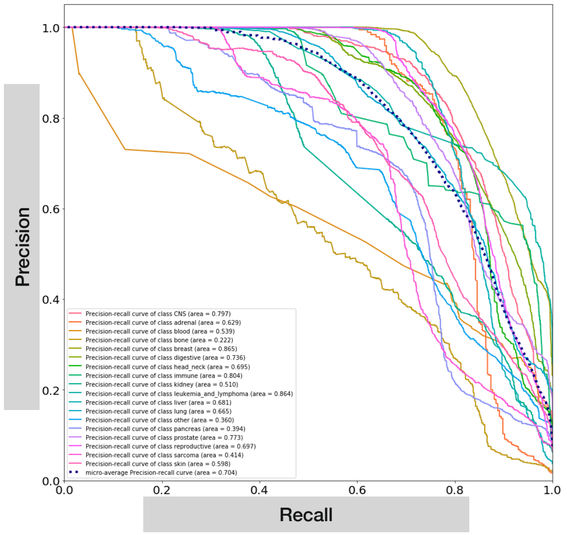
TRAINING SET

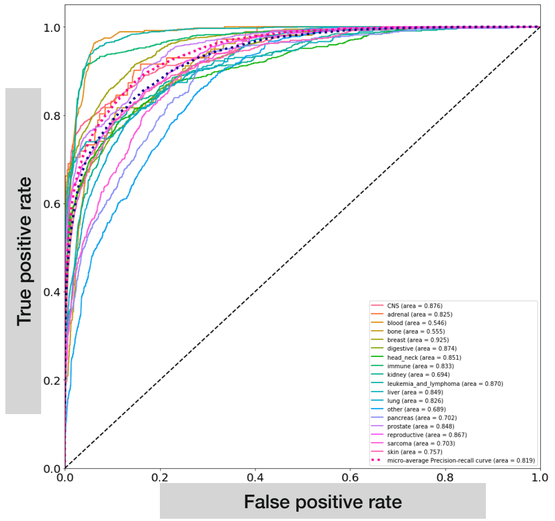


TEST SET



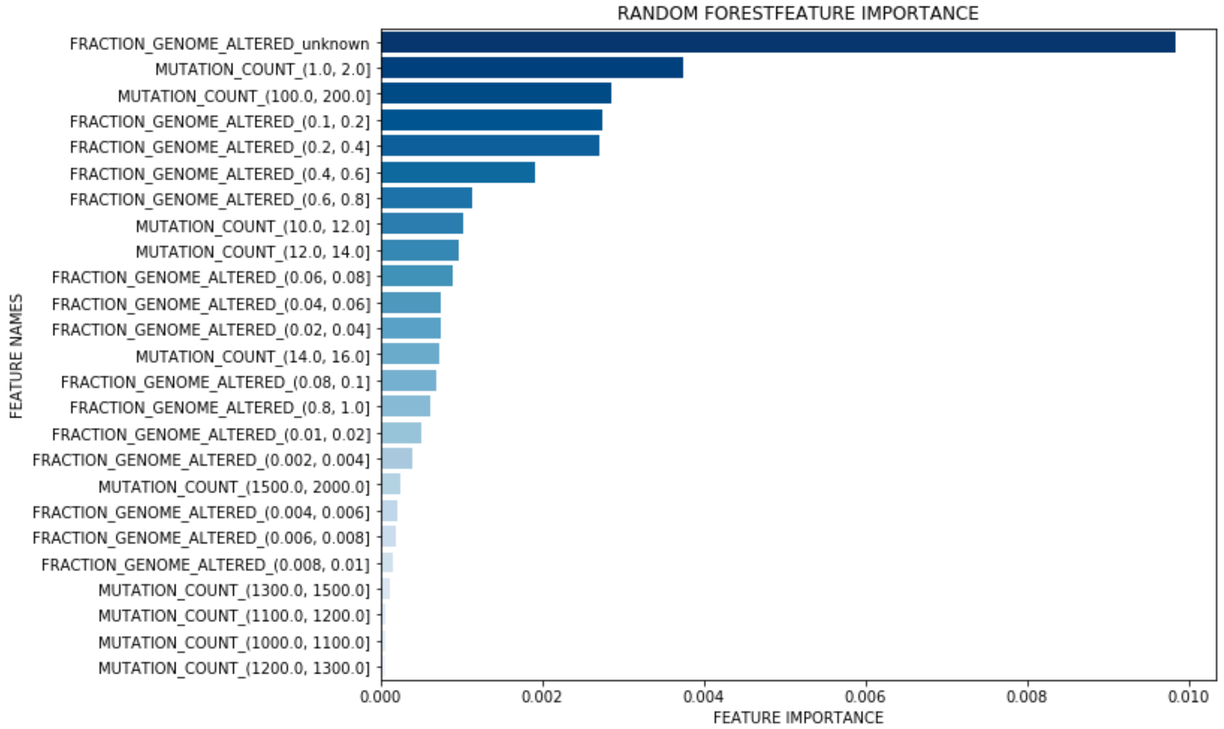
Model 2: Random forest - classification report for test set





	precision	recall	f1-score	support
CNS	0.77	0.73	0.75	1198
adrenal	0.91	0.58	0.71	71
blood	0.65	0.39	0.49	353
bone	0.56	0.15	0.23	158
breast	0.62	0.80	0.70	2154
digestive	0.76	0.62	0.68	1815
head_neck	0.78	0.59	0.67	1247
immune	0.51	0.91	0.65	834
kidney	0.67	0.34	0.45	635
leukemia_and_lymphoma	0.68	0.82	0.75	1425
liver	0.89	0.55	0.68	388
lung	0.46	0.74	0.56	1800
other	0.39	0.28	0.33	948
pancreas	0.68	0.32	0.44	598
prostate	0.65	0.73	0.69	1277
reproductive	0.77	0.62	0.68	788
sarcoma	0.64	0.24	0.35	487
skin	0.78	0.46	0.58	746
accuracy			0.63	16922
macro avg	0.68	0.55	0.58	16922
weighted avg	0.66	0.63	0.62	16922

Model 2: Random forest - top 25 features



Similarly, to the decision tree classifier, important features for the random forest model is **mutation count** and **fraction genome altered**.

Summary

Conclusion

This project was a preliminary investigation into the prediction of cancer types using clinical and genomic features. The models utilised here have successfully predicted cancer types 50% above the baseline score. The best performing model were the single decision tree and random forest models.

Limitations and future work

One of the limitations of this dataset is that there is limited data available for some of the cancer types resulting in an imbalance between the classes. Secondly, there is an inherent bias in feature reporting for solid and non-solid tumours. For example, it is not possible to report tumour size for a non-solid tumour. Therefore, such limitations should be addressed in future work.

LOGISTIC REGRESSION	DECISION TREE	RANDOM FOREST
MODEL 1 SCORE	**MODEL 2 SCORE**	**MODEL 3 SCORE**
0.62 Best score	0.62 Best score	0.62 Best score
0.62 Test score	0.63 Test score	0.63 Test score
0.64 Train score	0.68 Train score	0.74 Train score