The Sir Peter MacCallum Department of Oncology
CONFIRMATION REPORT

STUDENT NAME: Peter Donnelly	
MEETING DATE:	**VENUE**:
COMMITTEE:	
SUPERVISORS: Dr. Ken Doig and Dr. Thomas Conway	
MENTOR: Dr. Huiling Xu	
LENGTH OF CANDIDATURE TO DATE: 14 months	
PROJECT TITLE: Improving Classification of T-Cell Lymphomas By Applying Machine Learning to Heterogeneous Data Types	

Name of candidate: Peter Donnelly

Student ID Number: 1059982

Field of Research Codes: 111202, 111203 , 111207

Title of the thesis

"Improving Classification of T-Cell Lymphomas by Applying Machine Learning to Heterogeneous Data Types"

Abstract

It is proposed that a multi-modal neural network trained with matched heterogeneous data types, viz, Whole Slide Images (WSI) and gene expression data, may be an effective tool for automatically clasifying and sub-typing cancer tumours generally, and difficult to classify tumours such as certain kinds of T-Cell Lymphoma in particular.

The proposition will be tested by conducting experiments on matched WSI and rna-seq data using appropriate deep learning network models. The development of a suitable experiment system will be an key thesis deliverable.

Literature Review

Early and accurate cancer detection and classification is highy consequential: early detection allows treatment to commence before the disease has progressed too far; while accurate classification is the most important determinant of treatment programs and ultimately therefore, patient outcomes.

Cancer classifications taxonomies([1]) are the historical product of decades of effort on the part of oncolologists and medical scientists([2])([3]). Always contingent([4]), they are subject to change and refinement as and when new scientific evidence emerges which improves, refines or refines an existing classification([5]).

Until relatively recently, cancer classification was based on the visual evidence of histopathological analysis, but in the past two decades genomics evidence has come to play an important role in cancer classification([6])([7])([8])([9])([10])([11])([12]), in a way remarkably similar to the way DNA barcoding influences the morphology oriented Linnaean biological classification system([13]).

Deep Learning in Medicine

Deep Learning (DL)([14]) ([15]), a sub-category of Machine Learning named for its use of many layered Neural Networks, has made rapid progress in a wide variety of fields since about 2010, and is on the cusp of becoming clinically relevant in multiple medical fields([16])([17]), very much including Pathology([18])([19])and Oncology([20])([21])([22]).

Medical applications of Deep Learning can be divided into those focusing on research topics and those with direct clinical aims. Deep Learning has to date had much less impact in the clinical domain than it has in research domains (acknowledging that the former is inextricably and causally interwined with the latter). Notable exceptions include the diagnosis of skin cancer([23]) and eye disease, where the accuracy of Deep Learning is in some cases on a par with human diagnosticians([24])([25]).

Deep Learning in Pathology and Oncolgy

During the past five years in particular, researchers have applied Deep Learning in a variety of novel ways to multiple cancer artefacts including Whole Slide Images([26]) and all kinds of omics data, with diverse objectives and varying degrees of success([27]). These objectives include, but are not limited to: prediction of cancer driver genes([28]); identifying cancer associated signalling pathways(4); predicting gene expression levels from histopathology images([29]) ([30])([31]), recapitulating cell morphology from gene expression data([32]), improve variant calling([33]); cancer patient prognosis forecasting([34]); quantifying and segmenting Tumor Infiltrating Lymphocytes([35])([36]); and classifying cancer types and sub-types, of which mutliple examples will be discussed in subsequent sections.

Reflecting the relatively immature status of medical applications of DL, only a small number of papers specifically addressing methods and techniques([37])([38]).

Difficult to Classify Cancers

Pathologists find some kinds of cancers particularly challenging to classify([39]) ([40])([41]). Oftem only molecular testing can provide the additional information necessary to differentiate morphologically near identical subtypes. Lymphoma provides multiple examples of this: while many B-Cell Lymphomas and Non Hodgkins Lymphomas are straighforward to classify histologically, some of the large number of T-Cell Lymphoma sub-types are notoriously difficult to classify this way. As one example, the ALK-negative([42]) and ALK-positive([43])([44]) sub-types of Anaplastic Large Cell Lymphoma([45])([46]) have almost identical histomorphologies, and essentially cannot be differentiated visually([47])([48]). This is consequential, since the ALK-positive sub-type has a much better prognosis than the ALK-negative sub-type; the optimum treatment for each differs: the latter requring more aggressive treatment; and further ALK-negative sub-type is much more susceptible to relapse.

Deep Learning using Whole Slide Images

Many researchers have applied DL to histopathology images, inspired perhaps by the success of DL in image recognition, classification and segmentation in other domains; and facilitated by the widespread availability of open source implementations of effective highly DL algorithms, including VGG([49]), Resnet ([50]) and GoogLeNet/Inception ([51]), and machine learning frameworks, such as Tensorflow([52]), PyTorch([53]) or Caffe([54]).

A typical project of this type uses one or more of the above mentioned deep learning models to detect cancer and/or classify cancer type or subtypes from sets of very high resolution Whole Slide Images (WSI) scans of Haemotoxylin and Eosin (H&E) stained tissue samples or Core Needle Biopsy (CNB) samples. In most cases, detection and classification tackle focus on one or a small number of cancer types and sub-types, however some projects take the opposite tack and may attempt to classify 20 or more cancer types/sub-types)

Project scopes are frequently defined or constrained by the availability of suitable WSI data, very often making use of the TCGA([55]), or else take place in the context of a 'Grand Challenge'([56]) such as Camelyon17 ([57]), or ICIAR/BACH([58]) where curated data is made available to participants, and frequently left in place for future researchers to use.

Most but not by no means all projects use publicly available data. Most but not all use slide level truth labels. A small proportion engage pathologists to hand curate slides or tiles (portions of slides). While the latter approach should in principle lead to higher quality outputs, it is also costly and labour intensive, which may be why it is not seen more.

The following paragraphs survey contemporary applications of Deep Learning to Whole Slide Image data:

Campanella et al ([59]) (2019) classify four types of cancer from H&E stained WSI image samples using slide-level labels (n=17,661) with a ResNet 34 based DL system, claiming an accuracy of >98% for all types. Note that this is classification of cancer types, rather than the more challenging problem of classifying sub-types of a particular cancer.

Coudray et al ([60]) (2018) classify H&E stained WSI samples (n=1634) into either normal, lung adenocarcinoma (LUAD) and squamous cell carcinoma (LUSC), with an InceptionV3 based DL system, achieving an AUC of 0.97, which they claim is comparable to the performance of human pathologists. This experiment encompasses both cancer detection and sub-typing; albeit relatively simple sub-typing.

Brancati et al ([61]) (2019) classify three types of invasive ductal carcinomas: Chronic lymphocytic leukemia (CLL); mantle cell lymphoma (MCL) and Follicular Lymphoma (FL); from H&E stained breast tissue samples (n=162) using FusionNet (a type of Resdual Convolutional Neural Network); with a claimed accuracy of 98%. This experiment encompasses classifiying three distinct B-Cell Lymphoma sub-types.

Li et al ([62])(2020) (I) detected and segmented breast cancer tumour regions and (ii) classified and quantified tumour infiltrating lymphocytes in H&E stained breast cancer WSI samples (n=1015) using FusionNet, a type of Resdual Convolutional Neural Network.
This experminent encompasses TIL quantification and visualization in addition to cancer detection (but `does not include classification).

Pantanowitz et al ([63]) (2020) detected prostate cancer from H&E stained and core needle biopsy WSI images (n=1627) in the context of a clinical workflow using a Convolutional Neural Network, with a headline claimed AUC of 0.997, and including one case which human pathologists had missed.

Komura and Ishikawa([64]) (2018) is a rare and valuable example of a review paper on Machine Learning techniques for histopathological image analysis.

Deep Learning using Omics Data

Applying DL to omics data has been a popular target for researchers since about 2017. With multiple kiinds of omics data available, omics DL research objectives are unsurprisingly more diverse([65]) than is the case for WSI.
The following paragraphs survey contemporary applications of Deep Learning to omics data:

Gao et all ([66]) (2019) classified five different PAM50 ([67]) defined breast cancer sub-types using a Fully Connected Network trained with the TCGA rna-seq data andenriched by using the Molecular Signatures Database (MsigDB) ([68]),with a claimed accuracy of 80% (n=456).

Lee et all ([69]) (2020) sub-typed five different types of cancer using TCGA rna-seq data and the KEGG ([70]) pathway database using a Graph Convolutional Network (a specialized type of convolutional network designed to learn from input data that has the natural structure of a graph). They achieved classification accuracy of 91.5% for STAD (four sub-types) and a little less for the other three cancers. Additionally, their innovative use of a pathway database allowed them to add an element of explanation to the sub-typing.

Levy et all ([71]) (2020) predicted cancer type/sub-type, age and smoking status (inter alia) from essentially all TCGA methylation data using a comprehensive Deep Learning system ('MethylNet') which is based on Fully Connected Networks and Variational Autoencoders. It achieved 97% accuracy in classifying a mix of 32 cancer types and subtypes, and additionally 95% accuracy (n=1018) for PAM50 classification of breast cancer sub-types.

Luo et al (2) predict the top *n* driver colorectal and breast cancer driver genes using a one dimensional Convolutional Neural Network and compare these predictions with those of the literature, with an AUC of 0.984, claimed to be 15% higher than the best competing algorithm.

Al Mamun and Al Mamun (*[72]*) (2019) classified eight cancer types using Long Non-coding RNA Data (rna-seq) from UCSC Xena ([73]) TCGA sed four kinds of Neural Networks (Fully Connected, Convolutional, Long Short Term Memory and a Deep Autoencoder, achieving accuracies of 94-98%; and futher noted that lncRNA data appears to be superior in mRNA data for the classification task [a proposition my thesis will be readily able to test]

Multimodal Learning using both Image and Omics Data

'Multimodal Learning' here refers learning from the integration of than one kind of input data, for example; WSI data+rna-seq, or WSI data+methylation data, WSI data+rna-seq+methylation data. Specifically, outputs derive from *features* learned from the individual learning modes, rather than directly from input data.

There are currently few examples of Multimodal Learning in Pathology/Oncology, possibly because the research community tends to cleave between microscope/image oriented and genomics oriented. Notwithstanding, this small group includes some amibtious and highly innovative projects.

The following paragraphs survey contemporary applications of Multimodal Deep Learning to multiple data modalities:

Gundersen et al (*[74]*) *(2019)* use a combination of a Deep Convolutional Generative Network (DCGAN), Fully Connected Network (FCN) and Canonical Correlation Analysis (CCA) to extract mode specific and common mode information from image and rna-seq data using NIH's GTEx V6 dataset ([75]).The common mode information is used to recapitulate approximate versions of both the original images and expression data (primary objective); the trained model is also used to classify the 50 types of human tissue (n=2221) represented in the GTEx V6 dataset from common mode signal alone, and achieves p<=0.05 for 13 of the 50 tissue types. Although it has significantly different objectives, the Gundersen system besides the CCA component has many technical similarities with the system required for our experiments. The core learning engine of our system is an adapation and extension of the Gundersen system.

Carmichael et al (*[76]*) *(2019)* use the combination of a Deep Convolutional Network and a statistical technique analagous to CCA called AJIVE ([77]) to to extract, analyse and explain biologically meaningful features from multimodal data drawn from two sources: (i) image and gene expression data from the Carolina Breast Cancer Study (CBSC) (n=1191); and, (ii) four omics data types from the TCGA Breast Cancer data dataset (n=616).

Ash et al (*[78]*) *(2018)* use a combination of a Convolutional Autoencoders and Sparse Canonical Correlation Analysis to discover associations between sets of genes and physical cellular features/attributes; using image and gene expression data from two TCGA datasets and the GTEx dataset.

Cheerla et al ([79]) (2019) predict patients survial rates for 20 cancer types and 11,160 patients using TCGA data by using a customized Multimodal Neural Network incorporating CNNs, and FCNs ,to three data modalities: Whole Slide Images data, rna-seq data and clinical data with an overall AUC of 0.78.

Islam ([80]) et al perform PAM50 classification of breast cancer subtypes using two kinds of omics data, viz: rna-seq and copy number alteration (n=1925), using a Deep Convolutional Neural Network, achieving an AUC of 0.83.

Case for the Proposition of this Thesis

The research summarised in the preceding sections shows that the 'Deep Learning approach' has been successful in accomplishing tasks that centre on extracting complex patterns from biological data of arbitrary modality or combinations of modality, including cancer detection and classification.

Notwithstanding that this is a fast moving research area, the idea of using DL on a combination of both image and rna-seq data to classify and subtype cancers does not appear to have been pursued, to the best of our knowledge.

The proposition that multimodal approach might out-perform single mode DL in the classifying cancers is hardly a bold one.

The assumption that WSI image data and rna-seq data each contains at least *some* sub-type-relevant information that the other mode does not, is therefore a reasonable one, at least in principle; so we also reasonably expect that learning from the integrated combination of both ought yield a higher classification accuracy than either mode alone.

Further, the cell morphology observable in pathology images must embody an imperfect but *integrated synthesis* of the information contained in all underlying genonic information; whereas gene expression data by definition should be a 'more perfect' representation of a subset of genomic information: for example methylation data clearly contains information that rna-seq data does not, and that information clearly may affect cell morphology.

Proposition

That a multi-modal DL network trained with matched WSI and rna-seq data type is likely to be be an clinically effective tool for automatically sub-typing cancer tumours.

Beneficially, most of the methods and processes needed to test the current hypothesis are well defined by previous work. Likewise, many of their positive learnings can be taken on board; and hopefully also, pitfalls avoided.

Thesis Objectives

Aim 1

- Establish a baseline Experiment System using mainstream machine learning software, techniques and models and use this to test the hypothesis on ‘easy’ data, viz: publicly available data.
  - Develop the Experiment System
  - Locate, acquire and pre-process suitable matched experimental data
  - Test with WSI data
  - Test with gene expression data

Aim 2

- Demonstrate that a multi-modal DL network trained with matched WSI and rna-seq data type:
  - can consistently achieve better classification accuracy than any existing system which uses *either* WSI or rna-seq data
  - can consistently classify difficult to classify tumours
  - can consistently classify a dataset of difficult to classify T-Cell Lymphoma defined & provided by PMCC Oncologists.

Stretch objective for Aim 2

- Incorporate methylation beta value or copy number alteration (or both) as additional data modes

Aim 3

- Clinical Translation: Convert the Experiment System resulting from Aim 2 into a useful, usable and clinically relevant support tool for Curators and Pathologists.

Sources of Experimental Data

Although the objective is to perform experiments on PMCC’s own image and genomic data, it is unrealistic to assume that these will be available at an early date in the quantities required. It was therefore necessary to locate sources of public data.

After extensive searching, it became apparent that the only suitable public sources of matched WSI and omics data are (I) the NIH’s TCGA and (ii) NIH’s GTEx data repositories; and of these two, only NIH TCGA has the volumes and types of data required. Ostensibly there are many other public slide and omics data repositories world wide, but none have matched image and omics data; and further many ‘large data sets’ turn out to be highly heterogeneous, fragmented collections of small data sets which are often also unstructured or poorly curated.

More positively, the TCGA data repository, an intentional by-product of the decade long Pan Cancer project([81]) is *highly* suitable. It contains an enormous volume of uniformly structured and well curated cancer data, and a much smaller but nonetheless sufficient volume of case matched Whole Slide Image data.

Preliminary Experiments and Results

High Level Summary of Results

Early results are encouraging: using a TCGA Stomach and Adenocarcinoma cancer (STAD) dataset, which comprises 443intestinal adenocarcinoma - NOS cases spanning seven subtypes, and sample level truth labels, the system repeatedly achieves:

- >90% accuracy in classifying 229 WSI image samples
- >80% accuracy in classifying 479 rna-seq (60,483 mRNA + lnRNA) samples

As mentioned, asssuming that image data and rna-seq data each contain at least *some* subtype relevant information that the other does not, we expect that learning from the combination of both will yield an higher classification accuracy than either mode alone; ie. higher than 90% by some unknown amount.

Data

Experiments have been carried out on TCGA's DLBCL[1], SARC and STAD datasets, and with a small amount of PMCC T-Cell Lympohma data which has recently been made available to the project. Many hundreds of experimental runs have been conducted with TCGA STAD data

TCGA STAD is used as a reference dataset for the preliminary experiments because:

a) it contains enough subtypes (seven)[2] to make the task of classification a meaningful one
b) it contains enough pairs (219) of matched WSI and rna-seq data to make multimodal training feasible
c) it contains enough examples of five of the seven subtypes extraction of a resasonably balanced subset possible[3]
d) background tissue is relatively homogeneous, providing much less of a shortcut clue to cancer subtype compared to say, the TCGA sarcoma dataset

The seven STAD subtypes represented in the samples are:

- stomach adenocarcimoa – diffuse type

[1] The DLBCL dataset proved too small to be useful. The SARC dataset was large and has many subtypes, however it has been set aside because it was thought likely the systems' predictions would likely be influenced by background tissue morphology rather than tumour morphology, and no practical way of subtracting such an influence seemed possible.

[2] Two (stomach adenocarcinoma - signet ring type and intestinal adenocarcinoma - papillary type) are represented in very small numbers (13 and 8 respectively) and will not be used in future experiments

[3] 443 cases (229 + 479 samples) made up as follows: 72 x stomach adenocarcinoma diffuse type; 160 x stomach adenocarcinoma NOS; 21 x mucinous type intestinal adenocarcinoma; 82 x intestinal adenocarcinoma NOS; 79 x tubular type intestinal adenocarcinoma ; 13 x signet ring type stomach adenocarcinoma; 8 x papillary type intestinal adenocarcinoma; 3 x degenerate/unknown. The proportions of each class vary for the image and the rna-seq subset.

- stomach adenocarcinoma - NOS
- intestinal adenocarcinoma – mucinous type
- intestinal adenocarcinoma – tubular type
- intestinal adenocarcinoma – papillary type
- intestinal adenocarcinoma - NOS
- signet ring type adenocarcinoma

The Experiments

Key points about the preliminary experiments:

a) image experiments were carried out on 229 STAD WSI images, and while gene expression experiments were carried out on the 479 STAD rna-seq samples

b) Only absolute (rather than differential) gene expression values are used. Subsequent experiments will also use differential gene expression.

c) All 60,483 rna-seq values, comprising both protein coding messenger RNA (mRNA) and long non-coding messenger RNA (lnRNA), are used[4].

rna-seq Experiments

The preliminary rna-seq experiments:

a) use 479 samples; representing all seven subtypes

b) have 20% of samples of each type (image, rna-seq) exclusively held out for testing

c) use a 3 layer Fully Connected ('DENSE') network model reducing the dimensionality from 60,483 to the number of classes via a single hidden layer[5]

d) use Log10+1 data standardization

In reading the charts, most focus should be on cuves '*test loss*' and '*percent correct*'.

<u>Experiment r1</u>

Objective: ensure the model is genuinely learning

Outcome: the model is genuinely learning

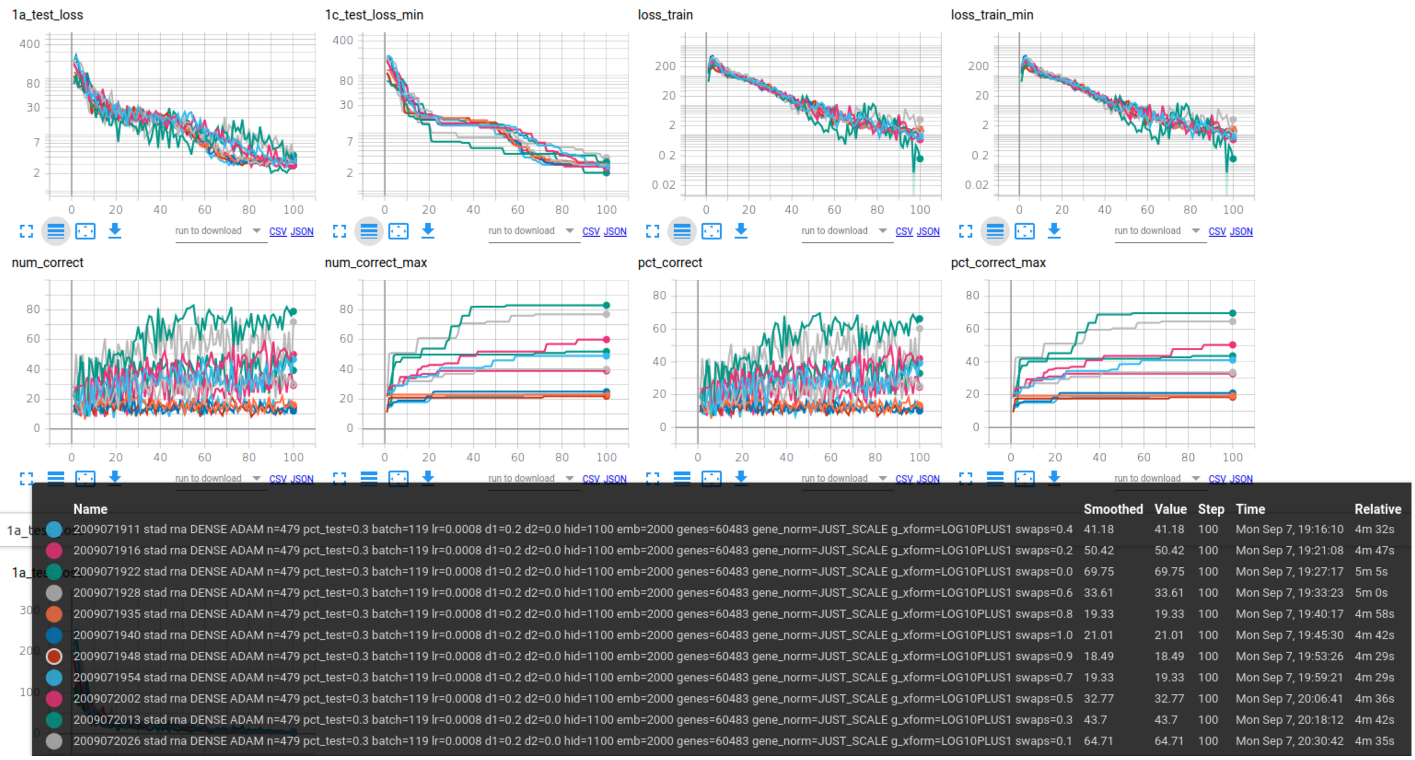
To ensure the network is actually learning from the data and not from hidden internal artfacts (wrong data, bugs), we randomly swap a proportion of the true class labels for class labels chosen at random.

[4] Subsequent experiments will use subsets of the 60,483 mRNA + lnRNA, including (i) just protein coding rna (ii) just genes represented in PMCC gene panels (iii) just genes which are known or suspected to be associated with a particular cancer, as defined in the lliterature or by PMCC scientists.

[5] Finding the dimensions of the hidden layer is the objective of one the experiments below – it turns out to be around 250. Deeper models were also tested, but at best these achieved results no better than the three layer model.

With all labels swapped, we’d expect predictions to achieve no better than chance. With *no* labels swapped, we’d expect the network to be genuinely learning to the extent it is able to learn given other parameters. As the proportion of labels swapped reduces, we expect predictions will be intermediate bewteen these two extremes. Each of the 11 runs in this experiment randomize a proportion of the samples’ truth labels, ranging from 100% to none of them, in -10% steps.

As can be seen in Figure   , the model is genuinely learning. The progression is most easiest to discern from the *num_correct_max* curves.



	Name	Smoothed	Value	Step	Time	Relative
●	2009071911 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.4	41.18	41.18	100	Mon Sep 7, 19:16:10	4m 32s
●	2009071916 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.2	50.42	50.42	100	Mon Sep 7, 19:21:08	4m 47s
●	2009071922 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	69.75	69.75	100	Mon Sep 7, 19:27:17	5m 5s
●	2009071928 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.6	33.61	33.61	100	Mon Sep 7, 19:33:23	5m 0s
●	2009071935 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.8	19.33	19.33	100	Mon Sep 7, 19:40:17	4m 58s
●	2009071940 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=1.0	21.01	21.01	100	Mon Sep 7, 19:45:30	4m 42s
●	2009071948 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.9	18.49	18.49	100	Mon Sep 7, 19:53:26	4m 29s
●	2009071954 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.7	19.33	19.33	100	Mon Sep 7, 19:59:21	4m 29s
●	2009072002 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.5	32.77	32.77	100	Mon Sep 7, 20:06:41	4m 36s
●	2009072013 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.3	43.7	43.7	100	Mon Sep 7, 20:18:12	4m 42s
●	2009072026 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.1	64.71	64.71	100	Mon Sep 7, 20:30:42	4m 35s

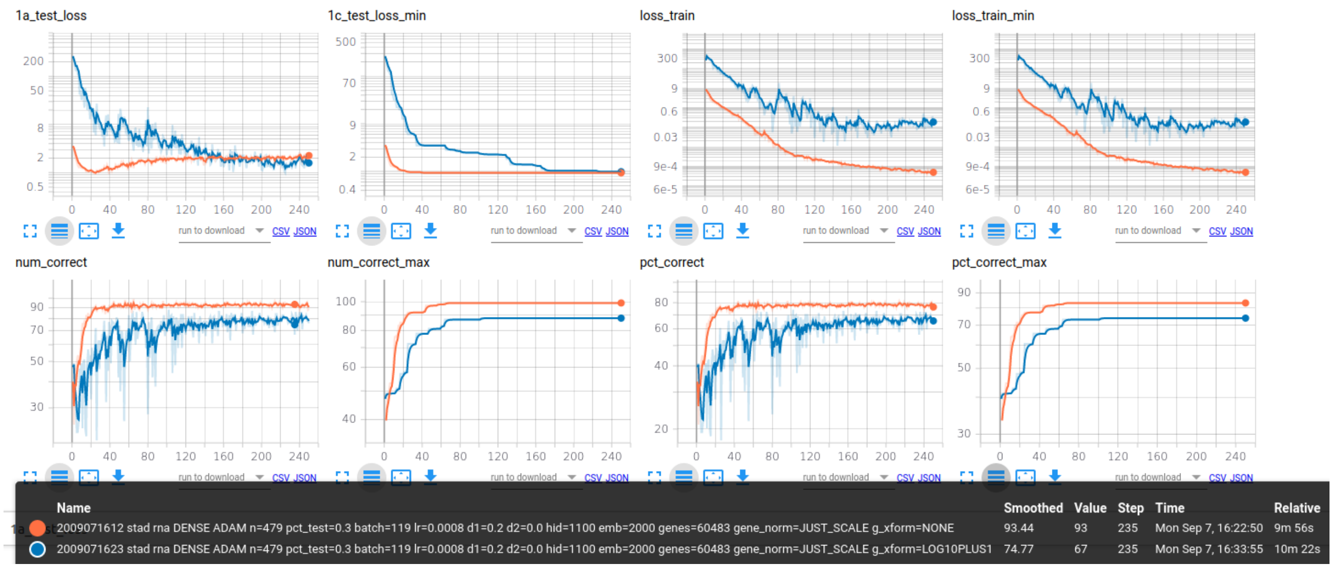
Experiment r2

Objective: Determine if scaling is required and if so what kind scaling should best be used.

Outcome: Log10+1 works the best. Log2+1 (not shown here) less so. Unscaled raw values works poorly.

Figure __ shows the dramatic impact of data scaling on the orders-of-magnitude-in-variance rna-seq data. Run 1 has no scaling, run 2 uses log10+1 scaling. All other parameters are identical and relatively well optimised.

With Log10+1 scaling, the lowest test loss occurs around epoch 39 where accuracy is 77%. Without scaling, the model takes much longer to get the test loss down, and achieves about 8% lower accuracy in the ‘steady state’.



	Name	Smoothed	Value	Step	Time	Relative
●	2009071612 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=NONE	93.44	93	235	Mon Sep 7, 16:22:50	9m 56s
○	2009071623 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1	74.77	67	235	Mon Sep 7, 16:33:55	10m 22s

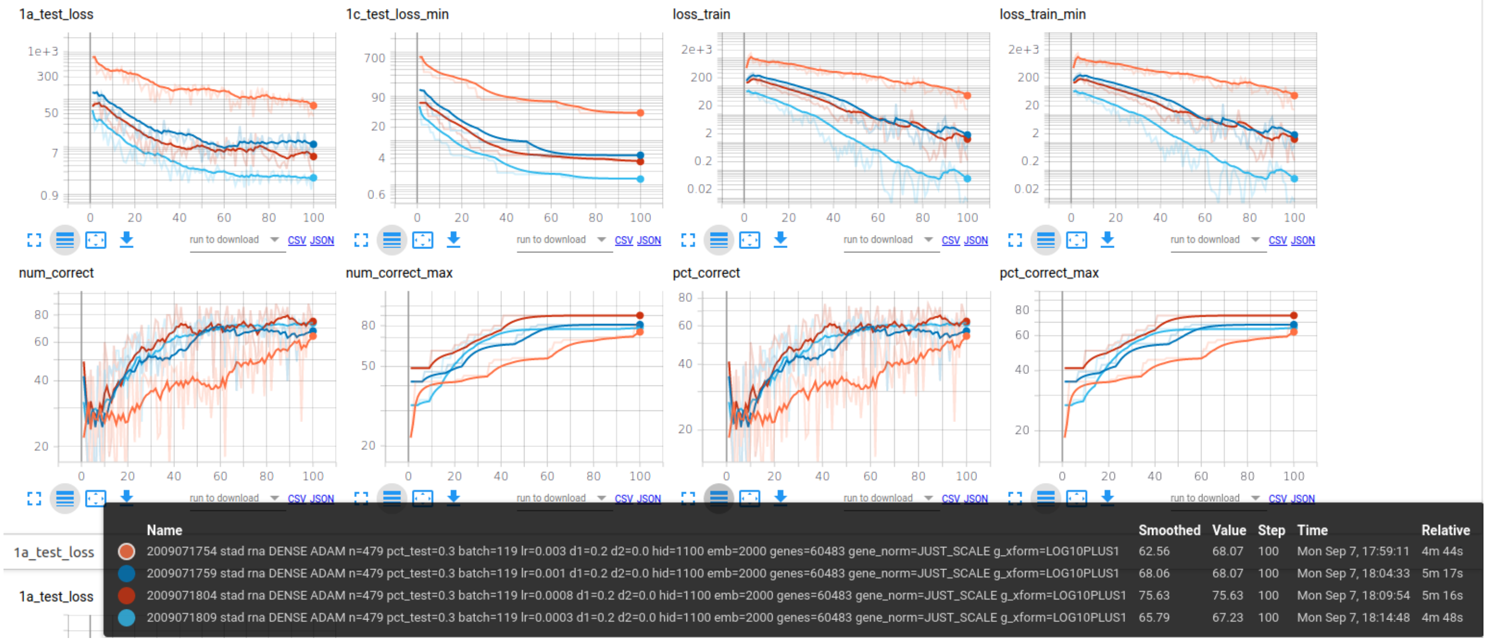
Experiment r3

Objective: Determine the best value for the learning rate hyperparameter

Outcome: The best performing value of those tested was ~0.0008

Learning rate refers to the proportion of the loss function gradient used to update network weights after each batch. The choice of learning rate makes a substantial difference to outcomes.

Running the model with a wide variety of learning rates (not just the ones shown here) reveal that the optimum learning rate is close to 0.0008 (.08%). In experiment 3, four learning rates have been tried, with all other parameters unchanged and reasonably optimised. The the deep red line corresponds to a learning rate of 0.0008.



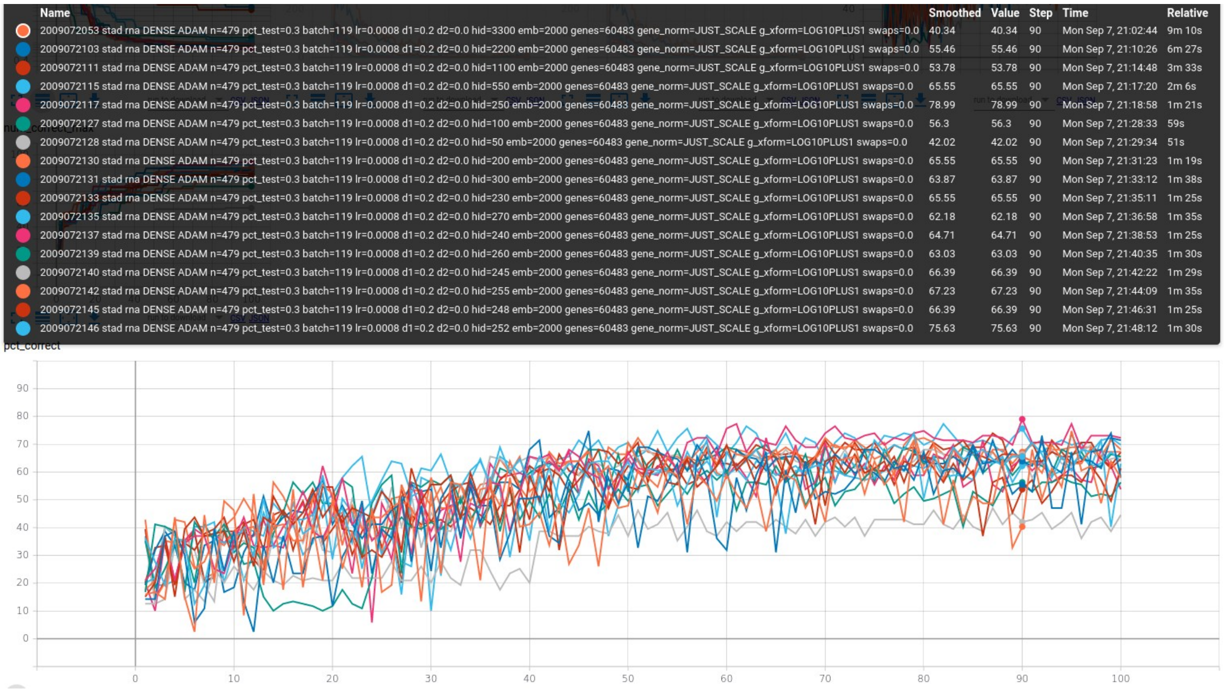
	Name	Smoothed	Value	Step	Time	Relative
●	2009071754 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.003 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1	62.56	68.07	100	Mon Sep 7, 17:59:11	4m 44s
●	2009071759 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.001 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1	68.06	68.07	100	Mon Sep 7, 18:04:33	5m 17s
●	2009071804 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1	75.63	75.63	100	Mon Sep 7, 18:09:54	5m 16s
●	2009071809 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0003 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1	65.79	67.23	100	Mon Sep 7, 18:14:48	4m 48s

Experiment r4

Objective: Determine the optimum number of neurons to use in the hidden layer of the Fully Connected Network

Outcome: The optimum number neurons to use in the hidden layer is ~250

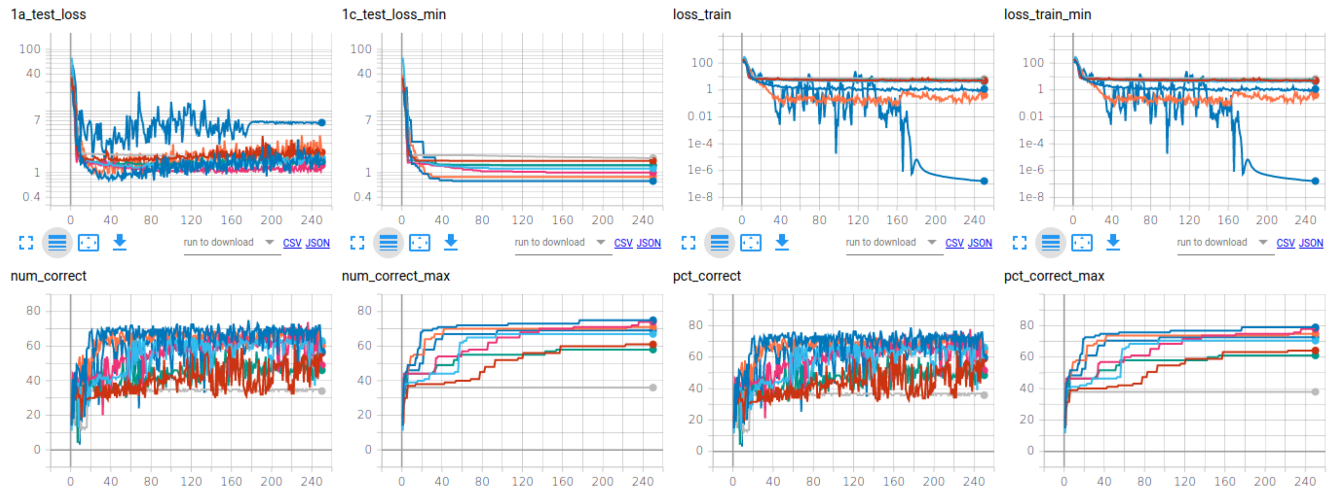
In this experiment the number neurons in the hidden layer is varied through the following set: (3300, 2200, 1100, 550, 250, 200, 300, 230, 270, 240, 260, 245, 255, 248, 252). The number of neurons corresponding to the least cost is 250 (pink curve), with lowest cost value occuring at epoch 91, where the test set achieved 79% accuracy – see image __ below.



	Name	Smoothed	Value	Step	Time	Relative
○	2009072053 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=3300 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	40.34	40.34	90	Mon Sep 7, 21:02:44	9m 10s
●	2009072103 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=2200 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	55.46	55.46	90	Mon Sep 7, 21:10:26	6m 27s
●	2009072111 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=1100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	53.78	53.78	90	Mon Sep 7, 21:14:48	3m 33s
●	2009072115 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=550 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	65.55	65.55	90	Mon Sep 7, 21:17:20	2m 6s
●	2009072117 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=250 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	78.99	78.99	90	Mon Sep 7, 21:18:58	1m 21s
●	2009072127 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=100 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	56.3	56.3	90	Mon Sep 7, 21:28:33	59s
●	2009072128 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=50 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	42.02	42.02	90	Mon Sep 7, 21:29:34	51s
●	2009072130 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=200 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	65.55	65.55	90	Mon Sep 7, 21:31:23	1m 19s
●	2009072131 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=300 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	63.87	63.87	90	Mon Sep 7, 21:33:12	1m 38s
●	2009072133 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=230 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	65.55	65.55	90	Mon Sep 7, 21:35:11	1m 25s
●	2009072135 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=270 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	62.18	62.18	90	Mon Sep 7, 21:36:58	1m 35s
●	2009072137 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=240 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	64.71	64.71	90	Mon Sep 7, 21:38:53	1m 25s
●	2009072139 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=260 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	63.03	63.03	90	Mon Sep 7, 21:40:35	1m 30s
●	2009072140 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=245 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	66.39	66.39	90	Mon Sep 7, 21:42:22	1m 29s
●	2009072142 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=255 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	67.23	67.23	90	Mon Sep 7, 21:44:09	1m 35s
●	2009072145 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=248 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	66.39	66.39	90	Mon Sep 7, 21:46:31	1m 25s
●	2009072146 stad rna DENSE ADAM n=479 pct_test=0.3 batch=119 lr=0.0008 d1=0.2 d2=0.0 hid=252 emb=2000 genes=60483 gene_norm=JUST_SCALE g_xform=LOG10PLUS1 swaps=0.0	75.63	75.63	90	Mon Sep 7, 21:48:12	1m 30s

Experiment r5

Objective: Determine a good value for the drop-out regularization hyperparameter



Outcome: Of those tested , the value of drop-out regulariztion with least cost was was 0.2

Droput regularization is a means of regularization whereby a proportion of randomly selected neurons in one layer is dropped after each iteration. It limits the ability of a network to overfit training data and improves generalization. In this experiment, dropout values of 0.0, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6 and 0.7 were imposed on the hiddentlayer; all other parameters unchanged and reasonably optimized.   The lower blue curve corresponds to 0.2. Lowest cost is at epoch 51 where percent correct = 73%.

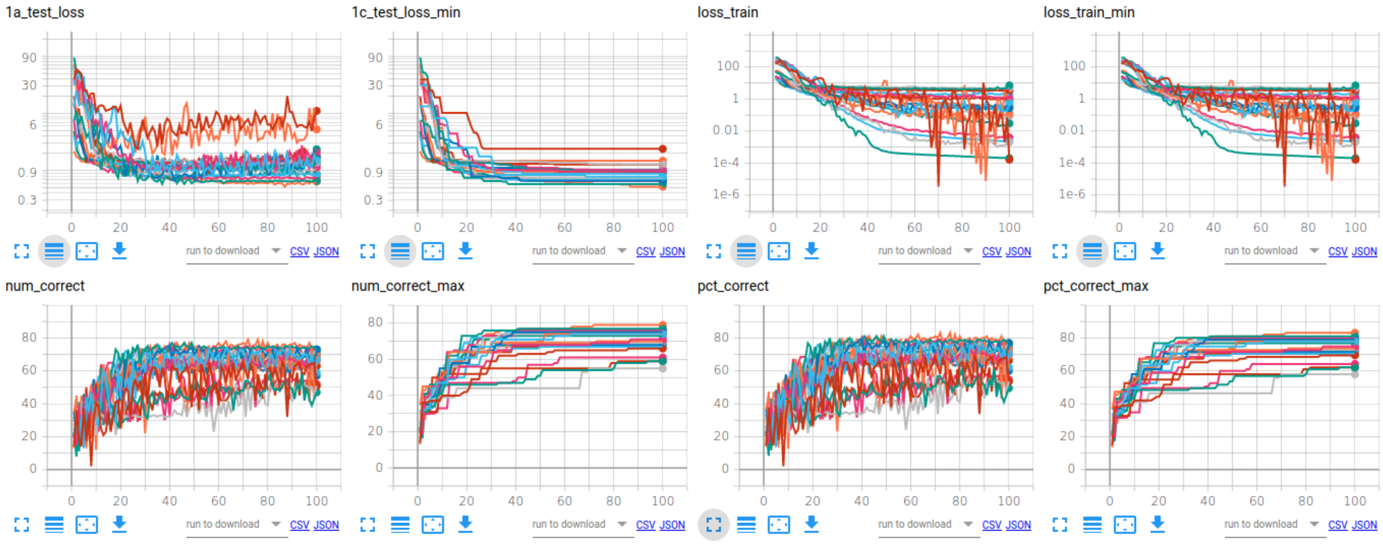
Experiment r6

Objective: Fine tune the hyperparameters deriving from Experiments r1 to r5.

Outcome: an improved result is obtained with drop-out = 0.4 and learning rate = . 0001,. Lowest cost occurs at epoch 87 where accuracy can be seen to have improved to 80%.

Using the nominally optimized hidden layer parameter (250 neurons), bracketing drop-out and learning rates around their nominally optimized values (0.2 and 0.0008 resp) to try and further improve the combination of hyperparameters.

The following set of drop-out values were used: 0.0 0.1 **0.2** 0.3 0.4 and the following set of learning rates (.001 .**0008** .0003 .0001 neurons) yielding an experiment job comprising 5x4=20 runs.



Experiment r7

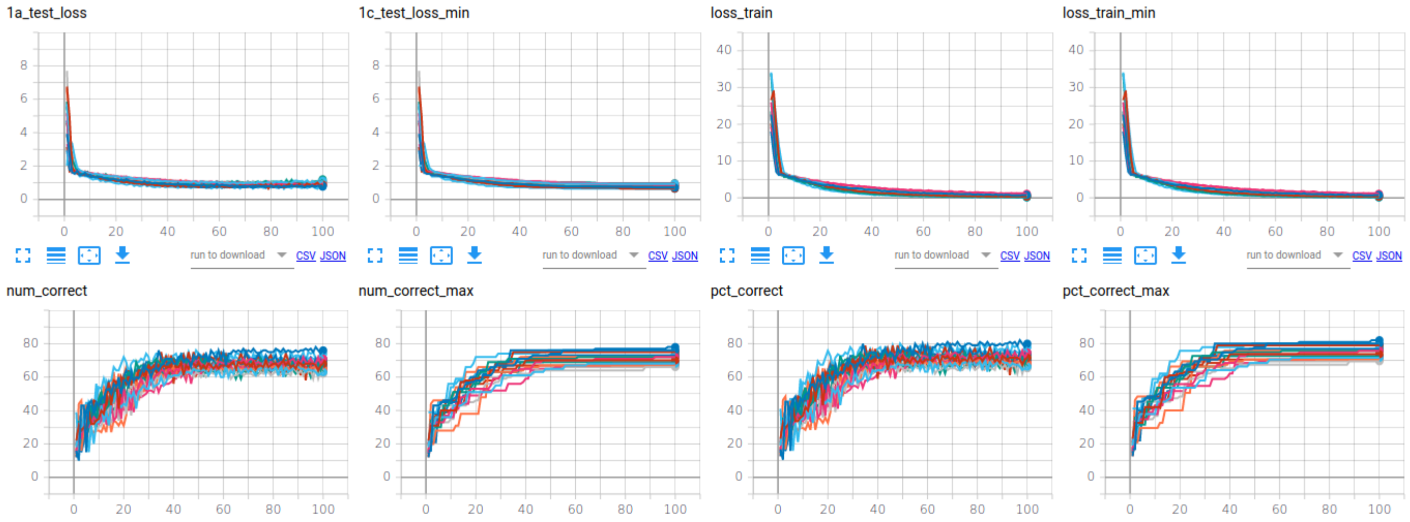
Objective: Verify that the the now notionally optimized model of produces reproducable results

Outcome: The optimized model produces reproducable results

Deep Learning based classification fits multidimensional curves to discrete data, which makes it a fundamentally stochastic procedure. While the held-out test set is sufficiently large (n=96) to ensure that results are unlikely to be influenced much by systematic or random factors, repeating the experiment with the same parameters plus small perturbations thereof would provide further comfort that systematic and random factors are not unduly influencing outcomes.

Thereore, a further set of experiments was conducted with drop-out values: (0.42 0.41 **0.40** 0.39 0.38) and learning rates (.00012 .00011 **.00010** .00009 .00008) yielding an job comprising 5x5=25 runs.

As can be seen from figure ___, the cost curves are very similar for all cases. The number/percent correct curves vary somewhat, but this is to be expected since test batches are drawn at random, so every test batch is unique. We also observe a slightly improved accuracy occuring with drop-out = 0.38 and learning rate = 0.00011, however this might not be 'real'



WSI Experiments

The preliminary image experiments use 229 samples; representing 7 subtypes; have 40% of samples held out for testing; and use a 11 layer VGG11 network model. A fixed number (e.g. 4000) of identically dimensioned (e.g. 64 x 64 pixles) tiles are extracted from random positions in each WSI sample, and these tiles are used for training (not the entire image).

Because the number of objects used for training is so large, the image experiments take much longer for images than for rna-seq; in the order of hours or small number of days. E.g. with 4,000 tiles per sample every epoch of training must process 229 x 4000 = 916,00 image tiles. This puts a practical limit on the number of experiments that can be conducted.

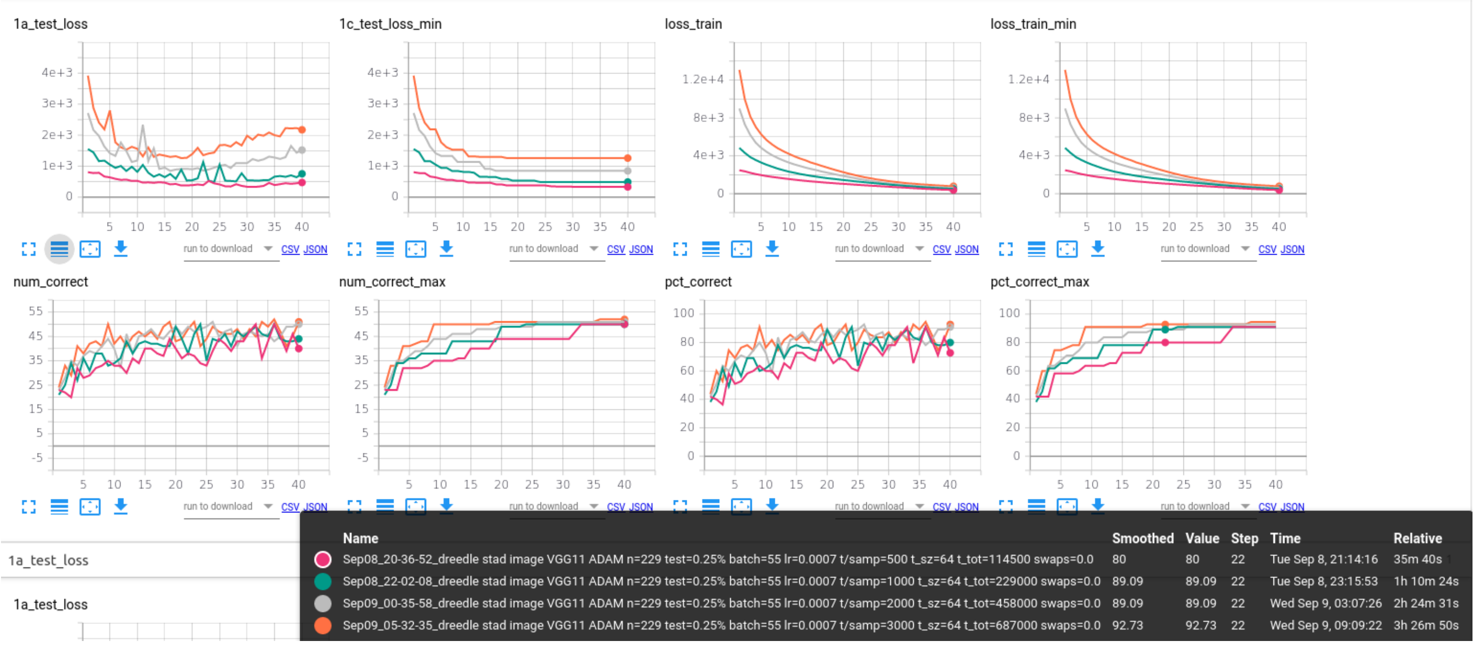
Experiment w1

Tiling is a computationally expensive task, so we wish to discover the minimum number of tiles per sample consistent with a good outcome. Since so many tiles in each sample will contain essentially the same information, it's not necessarily the case that more tiles will equate to higher classification accuracy.

A set of experiments was conducted with these tiles per WSI: (500 1000 2000 3000 4000).

Objective: determine the optimum number of tiles per sample to use

Outcome: As can be seen, the least cost curve (magenta) corresponds to 500 tiles per image, with the lowest cost value occuring at epoch 22, however this corresponds to an accuracy of only 69%. In fact, highest accuracy corresponds to the 3000 tiles per image curve (orange), closely followed by the 2000 columns per image curves (grey), with peak accuracies of 93% and 92% respectively. It's expected that the highest accuracy should occur on the least cost (500 tiles per sample), whereas the best accuracy achieved with the 500 tiles per sample case 89%, at epoch 33; so this requires further investigation. (Note that the size of the held-out test set is sufficiently large that the much higher accuracies obtained in the two cases mentioned is real).



	Name	Smoothed	Value	Step	Time	Relative
●	Sep08_20-36-52_dreedle stad image VGG11 ADAM n=229 test=0.25% batch=55 lr=0.0007 t/samp=500 t_sz=64 t_tot=114500 swaps=0.0	80	80	22	Tue Sep 8, 21:14:16	35m 40s
●	Sep08_22-02-08_dreedle stad image VGG11 ADAM n=229 test=0.25% batch=55 lr=0.0007 t/samp=1000 t_sz=64 t_tot=229000 swaps=0.0	89.09	89.09	22	Tue Sep 8, 23:15:53	1h 10m 24s
●	Sep09_00-35-58_dreedle stad image VGG11 ADAM n=229 test=0.25% batch=55 lr=0.0007 t/samp=2000 t_sz=64 t_tot=458000 swaps=0.0	89.09	89.09	22	Wed Sep 9, 03:07:26	2h 24m 31s
●	Sep09_05-32-35_dreedle stad image VGG11 ADAM n=229 test=0.25% batch=55 lr=0.0007 t/samp=3000 t_sz=64 t_tot=687000 swaps=0.0	92.73	92.73	22	Wed Sep 9, 09:09:22	3h 26m 50s

Experiment w2

Recalling that during training, only random batches of the held out test dataset are used for validation, we now more conclusively test the model by pass many more tiles through it.

Prodcess a patch of 12 x 12 tiles drawn from each image sample with highest accuracy model from Experiment W1. A total of 225[6] x 12 x 12 = 32,400 tiles (and therefore predictions) will be made.

Objective: Confirm the accuracy of the model with a 'full scale' test.

Outcome: ___% of the 32,400 predictions predicted the correct class.

Experiment w3

Predictions are made on a per-tile basis. We need to confirm that model classifications also make sense as a 'macro morphology' level.

[6] The batch size needs to be a perfect square to enable the patches to be square. Hence four samples arn't used.

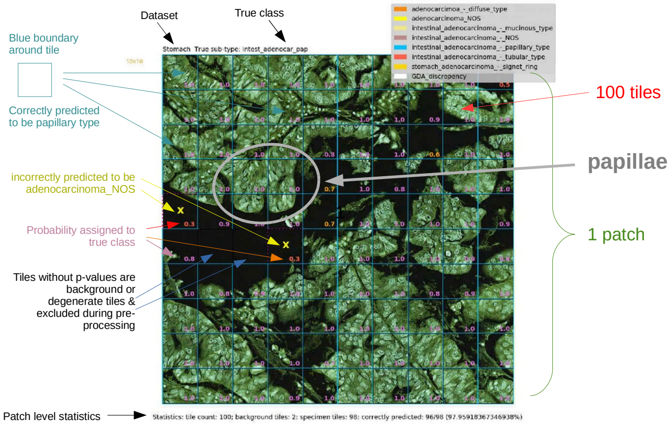
Objective: confirm that classifications produced by the model are ‘macro morphologically’ sensible.

Method: Extract 2D contiguous sets of tiles (a ‘patch’) from the samples and process using one of the high accuracy models from Experiment W1; annotating patches with applicable metadata including class predicted, true class and the probability the network assigned to the true class. Ask a PMCC histologist to review these.

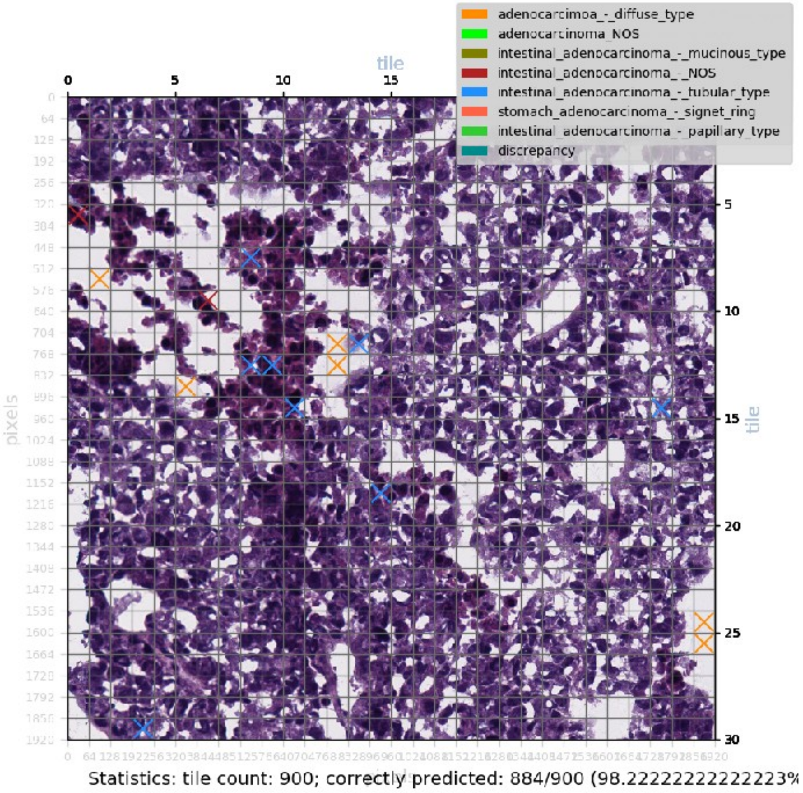
Outcome:

Each result is visualized in a manner similar to figure __.

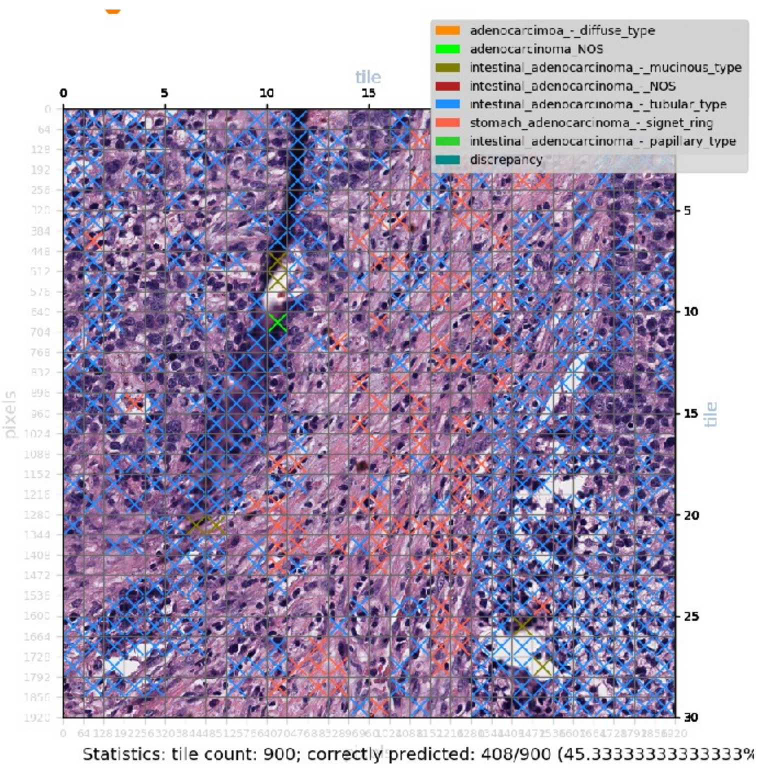
A subset of 12 of the 225, viz: 3 x confident correct predictions; 3 x unconfident correct predictions and 3 x wrong predictions will be provided to a pathology registrar to assess. In their assessment (…)


Dataset
True class
Blue boundary around tile
Correctly predicted to be papillary type
incorrectly predicted to be adenocarcinoma_NOS
Probability assigned to true class
Tiles without p-values are background or degenerate tiles & excluded during pre-processing
Patch level statistics
Stomach True sub-type: intest_adenocar_pap
10x10
adenocarcimoa_-_diffuse_type
adenocarcinoma_NOS
intestinal_adenocarcinoma_-_mucinous_type
intestinal_adenocarcinoma_-_NOS
intestinal_adenocarcinoma_-_papillary_type
intestinal_adenocarcinoma_-_tubular_type
stomach_adenocarcinoma_-_signet_ring
GDA_discrepency
100 tiles
papillae
1 patch
Statistics: tile count: 100; background tiles: 2; specimen tiles: 98; correctly predicted: 96/98 (97.95918367346938%)

Example of a Correct Prediction Made with High Confidence

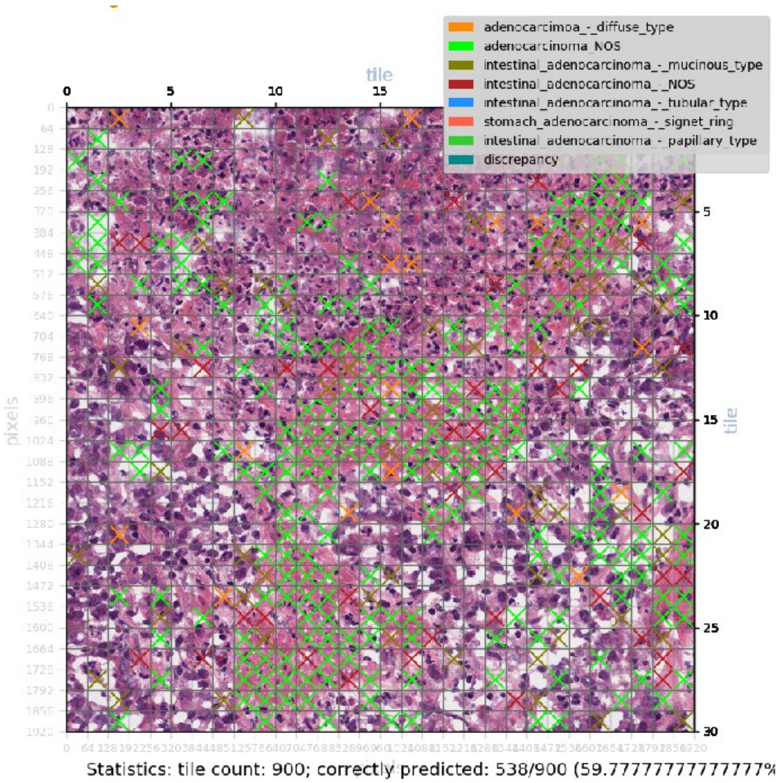

adenocarcimoa_-_diffuse_type
adenocarcinoma_NOS
intestinal_adenocarcinoma_-_mucinous_type
intestinal_adenocarcinoma_-_NOS
intestinal_adenocarcinoma_-_tubular_type
stomach_adenocarcinoma_-_signet_ring
intestinal_adenocarcinoma_-_papillary_type
discrepancy
tile
0
5
10
15
5
10
15
20
25
30
tile
pixels
Statistics: tile count: 900; correctly predicted: 884/900 (98.22222222222223%

Example of a Correct but Unconfident Prediction


adenocarcimoa_-_diffuse_type
adenocarcinoma_NOS
intestinal_adenocarcinoma_-_mucinous_type
intestinal_adenocarcinoma_-_NOS
intestinal_adenocarcinoma_-_tubular_type
stomach_adenocarcinoma_-_signet_ring
intestinal_adenocarcinoma_-_papillary_type
discrepancy
tile
0
5
10
15
5
10
15
20
25
30
tile
pixels

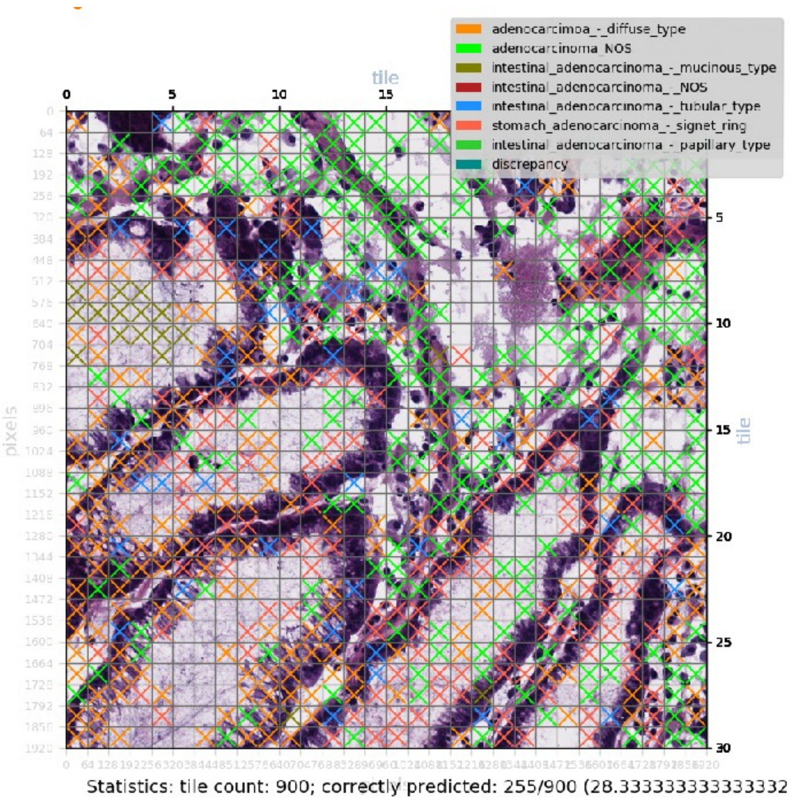
Statistics: tile count: 900; correctly predicted: 408/900 (45.33333333333333%

Example of a Wrong Prediction


adenocarcimoa_-_diffuse_type
adenocarcinoma_NOS
intestinal_adenocarcinoma_-_mucinous_type
intestinal_adenocarcinoma_-_NOS
intestinal_adenocarcinoma_-_tubular_type
stomach_adenocarcinoma_-_signet_ring
intestinal_adenocarcinoma_-_papillary_type
discrepancy
tile
0
5
10
15
5
10
15
20
25
30
tile
pixels

Statistics: tile count: 900; correctly predicted: 538/900 (59.77777777777777%

Example of a Badly Wrong Prediction


adenocarcimoa_-_diffuse_type
adenocarcinoma_NOS
intestinal_adenocarcinoma_-_mucinous_type
intestinal_adenocarcinoma_-_NOS
intestinal_adenocarcinoma_-_tubular_type
stomach_adenocarcinoma_-_signet_ring
intestinal_adenocarcinoma_-_papillary_type
discrepancy
tile
0
5
10
15
5
10
15
20
25
30
tile
pixels
0
64
128
192
256
320
384
448
512
576
640
704
768
832
896
960
1024
1088
1152
1216
1280
1344
1408
1472
1536
1600
1664
1728
1792
1856
1920
Statistics: tile count: 900; correctly predicted: 255/900 (28.333333333333332

Summary of Progress to Date

Summary

Aim 1 is largely complete.

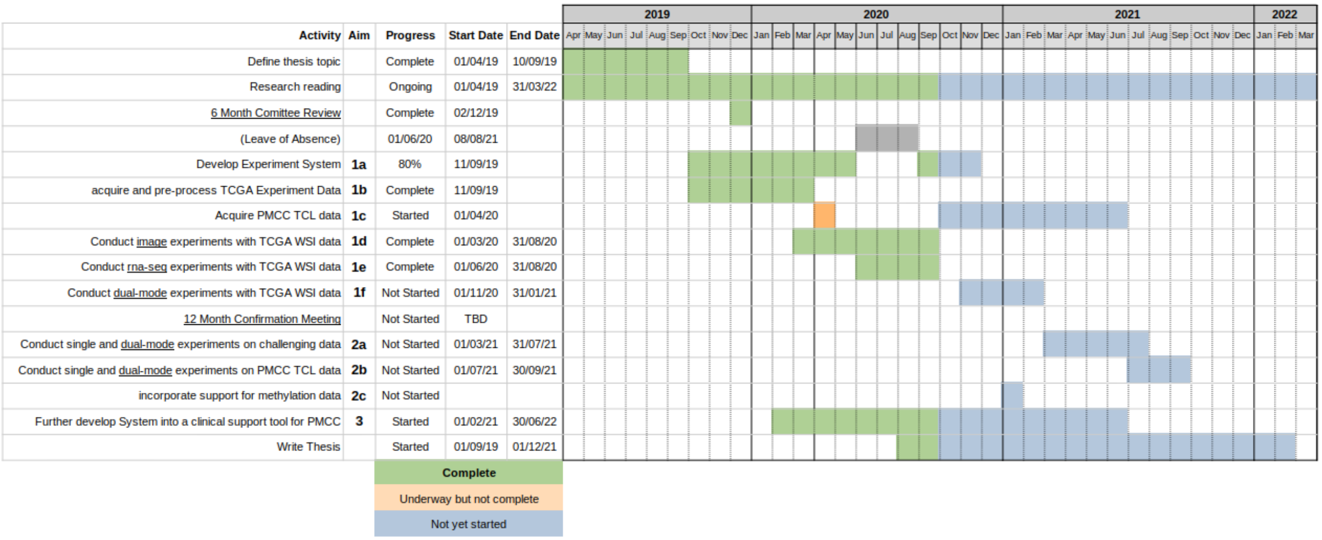
- Development of the Experiment System (Aim 1a) is approximately 90% complete. In its current state, the system is capable of detecting and classifying/sub-typing cancers from either whole slide image data or gene expression data (but not both together).
- Testing with TCGA WSI data (Aim 1b) is complete
- Testing with TCGA gene expression data (Aim 1b) is complete.
- Testing with dual mode WSI + rna-seq data has not yet commenced

Aim 2, which will see the system classify challenging cancer types, has a dependency on the availability of internal PMCC samples, for which a firm timeline is not presently available (a small batch of 10 T-Cell Lymphoma samples has been provided and used to validate that the system can handle CNB samples).

The Aim 2 stretch objective (2c) will commence in late 2020 or early 2021. It will be relatively straightforward since methylation data is structurally similar to rna-seq data.

Aim 3 commenced at the same time as Aim 1, since the system has been designed and architected to be a clinically robust support tool from the outset. Aim 3 but will extend for a further 4 months beyond the end of stage 2, and complete in mid 2021.

Research Plan



					2019									2020												2021												2022		
Activity	**Aim**	**Progress**	**Start Date**	**End Date**	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sep	Oct	Nov	Dec	Jan	Feb	Mar
Define thesis topic		Complete	01/04/19	10/09/19	🟩	🟩	🟩	🟩	🟩	🟩																														
Research reading		Ongoing	01/04/19	31/03/22	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦
6 Month Comittee Review		Complete	02/12/19										🟩																											
(Leave of Absence)		01/06/20	08/08/21																▒	▒	▒																			
Develop Experiment System	1a	80%	11/09/19								🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩				🟩	🟦	🟦																
acquire and pre-process TCGA Experiment Data	1b	Complete	11/09/19								🟩	🟩	🟩	🟩	🟩	🟩																								
Acquire PMCC TCL data	1c	Started	01/04/20														🟧						🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦									
Conduct image experiments with TCGA WSI data	1d	Complete	01/03/20	31/08/20												🟩	🟩	🟩	🟩	🟩	🟩	🟩																		
Conduct rna-seq experiments with TCGA WSI data	1e	Complete	01/06/20	31/08/20															🟩	🟩	🟩	🟩																		
Conduct dual-mode experiments with TCGA WSI data	1f	Not Started	01/11/20	31/01/21																				🟦	🟦	🟦	🟦													
12 Month Confirmation Meeting		Not Started	TBD																																					
Conduct single and dual-mode experiments on challenging data	2a	Not Started	01/03/21	31/07/21																								🟦	🟦	🟦	🟦	🟦								
Conduct single and dual-mode experiments on PMCC TCL data	2b	Not Started	01/07/21	30/09/21																												🟦	🟦	🟦						
incorporate support for methylation data	2c	Not Started																								🟦														
Further develop System into a clinical support tool for PMCC	3	Started	01/02/21	30/06/22											🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟩	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦									
Write Thesis		Started	01/09/19	01/12/21																	🟩	🟩	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	🟦	

🟩 **Complete**

🟧 Underway but not complete

🟦 Not yet started

Research Outputs

- Jun 2020 presentation to Bioinformatics Seminar: “Notes on Developing a Multimodal Deep Learning Network to Classify Cancers using Whole Slide Image and Gene Expression Data
- Jan 2020 presentation to Bioinformatics Journal Club on paper: “End-to-end training of deep probabilistic CCA forjoint modeling of paired biomedical observations”
- Jun 2019 presentation to PeterMac Machine Learning Seminar on paper: “Text-mining clinically relevant cancer biomarkers for curation into the CIViC database”
- Participation in VCCC Molecular Tumour Boards
- Co-organizer of / participation in Bioinformatics Journal Club
- Participation in PMCC Bioinformatics Seminars

Record of attendance

- Attendance at 2019 Methods in Cancer seminar series
- Post graduate research induction and orientation
- Completion of PMCC Occupational Health And Safety (BP&C) training
- Completion of PMCC Laboratory Safety Expiry training
- Completion of PMCC Emergency Procedures training

Appendix: Development of a Suitable Experiment System

Learnings from the Literature Review

These derive from a synthesis of the entire literature review rather than any individual paper; in some cases in combination with the candidate's technical computing know how.

High level Learnings (strategic / system / methods / techniques)

a) Use proven open source learning algorithms supported and maintained by the computer science community. Avoid 'novel' and 'bespoke' algorithms.

b) TCGA is the by far the best (and essentially the only) public source of a suitable volume of labeled multimodal case data.

c) Use Multiple Instance Learning and slide level labels rather than manual tile curation.

d) Data acquisition from TCGA is likely to be repetitive and time consuming, so it should be automated

e) Data pre-processing is likely to be complex and time consuming, so it should be automated

f) The native dimensionality of the TCGA rna-seq data is so high the a suitable means of dimensionality reduction may be required (e.g. Autoencoder or)

g) The TCGA rna-seq data includes gene expression values for both long non-coding RNA (lnRNA) in addition to protein coding mRNA. The experiment system should allow for the use of either of these sets, and additionally the subset of mRNA corresponding to PMCC gene panel(s) as well as subsets of genes which are already known to be implicated in particular cancers.

h) We hypothesize that the image branch will likely benefit from the ability to accept image tiles with multiple levels of magnification, so the experiment system should cater for this

i) Digital Stain Normalization may be required to compensate for batch variations in staining, so the experiment system should cater for this

j) The experiment system must be able to cater for both H&E stained tissue slides and aspirate smears

k) It is unknown whether data augmentation (e.g. hue variation) will be beneficial, but the experiment system should assume that it will be, and cater for it.

l) We suspect that Transfer Learning (eg. from ImageNet, which some of the projects use) is likely to be of zero or marginal benefit, so it need not be catered for in the first instance

Low Level (software, algorithm and choices of parameters)

a) Use Python/Pytorch rather than Tensorflow/Keras or Caffe as the DL engine

b) The ubiquity of Leica/Aperio slide scanning systems means that the system must cater for SVS image format in addition to the TIF image format

c) The image analysis branch should support at least the following network models: VGGnn, INCEPTION and RESNET

d) The rna analysis branch requires only the Fully Connected Network model (/ANN / 'Multi-Layer Perceptron'

e) The optimizer of choice is likely to be ADAM ('Adaptive Moment Estimation'), although others should be tried (eg. ADAGRAD, SGD …)

f) The classification loss function will certainly be Cross Entropy

Experiment System

Testing the hypothesis requires the development of a flexible experimental system capable of being trained multimodally from matched whole slide image and gene expression data. The development of such a system is an key deliverable of the thesis.

A single user defined experiment 'job' defines & runs multiple experiments, each with selectable combination of parameters/hyperparameters. Job level experiment definition has already shown itself to be a critical tool for experiment efficiency, consistency and repeatability.

The Experiment System supports, via PyTorch, CUDA(82)/CUDNN(83) compliant GPUs. The system does not require GPUs to operate, but with the size and volumes of data being processed, GPUs are inevitably necessary. If more than one GPU is available, the system will take vantage of all GPUs and perform parallel processing (currently only implemented for Autoencoder mode, but will likely be implemented for image and rna-seq mode in due course). Additionally, the Data Analysis subsystem uses the CuPy(82) to take advantage of GPU processors and memory pooling when performing correlation and covariance analysis.

The system builds on and extends open source software made available by other researchers, but the greater proportion of it is the original work of the candidate.

Subsystems and Capabilities:

The experiment system comprises the following following functional subsystems/capabilities:

1 <u>Biodata Acquisition Subsystem</u>

Automatically acquires data from NIH GDC repository (via its RESTful(82)) API in accordance with user provided filters; for example the following command will download not more than 2000 matched STAD(83) image and rna-seq files then save them in the structured manner expected by the Pre-processing subsystem.

```
./gdc_fetch.py --case_filter="filters/TCGA-STAD_case_filter"
--file_filter="filters/GLOBAL_file_filter_UQ" --max_cases=2000 --max_files=3  --
global_max_downloads=2000 --output_dir=stad
```

2 Pre-processing Subsystem

- Comprising the following capabilities:

  a) rapid identification and removal of background, degenerate and low contrast tiles using stochastic algorithms

  b) extraction of tiles from each high resolution WSI image sample in accordance with user defined quality parameters

  c) (optional) stain colour normalization(82) ('reinhard(83)' or 'spcn(84)')

  d) (optional) gene set selection using ENSEMBL(82) or PMCC 'cancer genes of interest' or user defined custom gene set

  e) (optional) normalization of images and rna-seq data (JUST_SCALE/GAUSSIAN; LOG10PLUS1/ LOG2PLUS1

  f) final generation of Pytorch ready input dataset (including optional generation of matched image+rna-seq dataset)

3 Experiment Management Subsystem

- Takes the output of the Pre-processing subsystem and runs experiments on it according to user defined parameters
- Four modes: image, rna, image_rna and autoencoder
- Experiment job definition: arbitrary combinations of the following parameters (as an unrealistic but valid example, in rna-seq mode, the underlined parameters would launch a job comprising 2x2x2x2x2x3x2x3x2= 1,152 separate experiments, and the all experiments would then run sequentially without further user supervision)

  a) DL models (e.g. “DENSE CONV1D”)

  b) optimization algorithms (e.g. “ADAM SGD”)

  c) tiles per slide (e.g. “500 1000 2000”)

  d) learning rates(s) (e.g. “.007 .001”)

  e) number of samples to use (e.g. “100 200”)

  f) mini-batch size(s) (e.g. “64 128”)

  g) hidden layer neurons for Fully Connected models (e.g. “10000 90000 8000”)

  h) topology definition for deep FCN models (one topology per job) (e.g. “3000 2000”)

i) optional dropout regularization on layer 1 of DENSE models (e.g. “0.2 0.3”)

j) optional dropout regularization on layer 2 of DENSE models (e.g. “0.2 0.3”)

k) optional stain normalizations (e.g. “NONE REINHARD SPCN”)

l) optional data transformations (rotations, flips, greyscale …)

m) optional data normalizations (e.g. “NONE JUST SCALE GAUSSIAN”)

n) optional data standardizations (e.g. “NONE LOG10PLUS1”)

o) optional noise and randomization data (for testing purposes: jitter, class randomization)

- Learning models

a) 3 x models & variants for image data ('VGG11/13/16/19', 'RESNET', INCEPTION V3')

b) 2 x models and variants for gene data (DENSE, 1D CNN)

c) 2 x models and variants for autoencoding (FC, DEEP, Variational)

4 Output Instrumentation Subsystem

a) real-time graphical view of results during training of training/test loss loss and accuracy curves

b) real-time console view of experiments as they are running

5 Output Visualization Subsystem

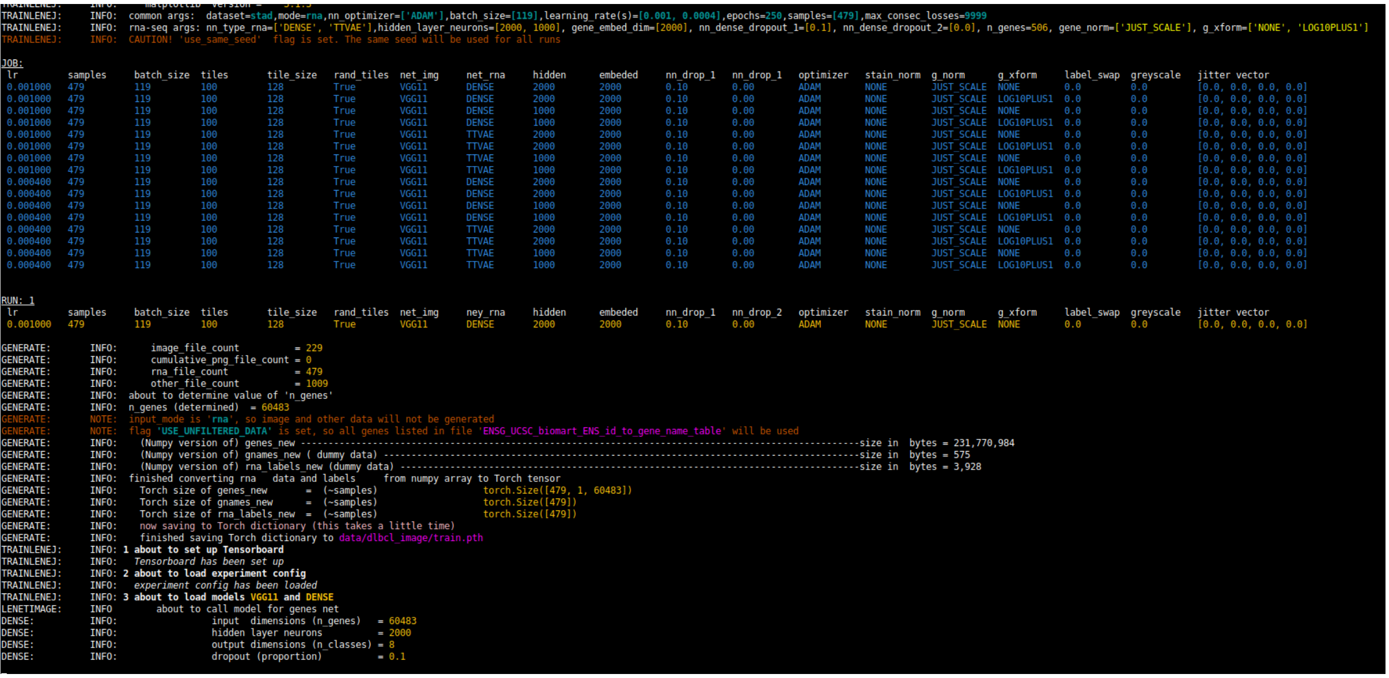
Real-time viewing modes are as follows:

a) test image tiles annotated with predicted labels vs ground truth labels (per batch)

b) test patches annotated with predictions and probabilities for each class

c) heatmaps showing confidence of correct/incorrect predictions

d) expression level correlation and covariance between genes for chosen gene set



```
TRAINLENEJ:     INFO:   common args:  dataset=stad,mode=rna,nn_optimizer=['ADAM'],batch_size=[119],learning_rate(s)=[0.001, 0.0004],epochs=250,samples=[479],max_consec_losses=9999
TRAINLENEJ:     INFO:   rna-seq args: nn_type_rna=['DENSE', 'TTVAE'],hidden_layer_neurons=[2000, 1000], gene_embed_dim=[2000], nn_dense_dropout_1=[0.1], nn_dense_dropout_2=[0.0], n_genes=506, gene_norm=['JUST_SCALE'], g_xform=['NONE', 'LOG10PLUS1']
TRAINLENEJ:     INFO:   CAUTION! 'use_same_seed'  flag is set. The same seed will be used for all runs

JOB:
 lr         samples    batch_size  tiles      tile_size   rand_tiles  net_img    net_rna    hidden     embeded    nn_drop_1  nn_drop_1  optimizer  stain_norm  g_norm      g_xform     label_swap  greyscale   jitter vector
 0.001000   479        119         100        128         True        VGG11      DENSE      2000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  NONE        0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.001000   479        119         100        128         True        VGG11      DENSE      2000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  LOG10PLUS1  0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.001000   479        119         100        128         True        VGG11      DENSE      1000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  NONE        0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.001000   479        119         100        128         True        VGG11      DENSE      1000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  LOG10PLUS1  0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.001000   479        119         100        128         True        VGG11      TTVAE      2000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  NONE        0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.001000   479        119         100        128         True        VGG11      TTVAE      2000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  LOG10PLUS1  0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.001000   479        119         100        128         True        VGG11      TTVAE      1000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  NONE        0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.001000   479        119         100        128         True        VGG11      TTVAE      1000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  LOG10PLUS1  0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.000400   479        119         100        128         True        VGG11      DENSE      2000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  NONE        0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.000400   479        119         100        128         True        VGG11      DENSE      2000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  LOG10PLUS1  0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.000400   479        119         100        128         True        VGG11      DENSE      1000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  NONE        0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.000400   479        119         100        128         True        VGG11      DENSE      1000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  LOG10PLUS1  0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.000400   479        119         100        128         True        VGG11      TTVAE      2000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  NONE        0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.000400   479        119         100        128         True        VGG11      TTVAE      2000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  LOG10PLUS1  0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.000400   479        119         100        128         True        VGG11      TTVAE      1000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  NONE        0.0         0.0         [0.0, 0.0, 0.0, 0.0]
 0.000400   479        119         100        128         True        VGG11      TTVAE      1000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  LOG10PLUS1  0.0         0.0         [0.0, 0.0, 0.0, 0.0]


RUN: 1
 lr         samples    batch_size  tiles      tile_size   rand_tiles  net_img    ney_rna    hidden     embeded    nn_drop_1  nn_drop_2  optimizer  stain_norm  g_norm      g_xform     label_swap  greyscale   jitter vector
 0.001000   479        119         100        128         True        VGG11      DENSE      2000       2000       0.10       0.00       ADAM       NONE        JUST_SCALE  NONE        0.0         0.0         [0.0, 0.0, 0.0, 0.0]

GENERATE:       INFO:      image_file_count          = 229
GENERATE:       INFO:      cumulative_png_file_count = 0
GENERATE:       INFO:      rna_file_count            = 479
GENERATE:       INFO:      other_file_count          = 1009
GENERATE:       INFO:   about to determine value of 'n_genes'
GENERATE:       INFO:   n_genes (determined)  = 60483
GENERATE:       NOTE:   input_mode is 'rna', so image and other data will not be generated
GENERATE:       NOTE:   flag 'USE_UNFILTERED_DATA' is set, so all genes listed in file 'ENSG_UCSC_biomart_ENS_id_to_gene_name_table' will be used
GENERATE:       INFO:     (Numpy version of) genes_new -------------------------------------------------------------------------------------------------size in  bytes = 231,770,984
GENERATE:       INFO:     (Numpy version of) gnames_new ( dummy data) ---------------------------------------------------------------------------------size in  bytes = 575
GENERATE:       INFO:     (Numpy version of) rna_labels_new (dummy data) -------------------------------------------------------------------------------size in  bytes = 3,928
GENERATE:       INFO:   finished converting rna   data and labels     from numpy array to Torch tensor
GENERATE:       INFO:     Torch size of genes_new      =  (~samples)                  torch.Size([479, 1, 60483])
GENERATE:       INFO:     Torch size of gnames_new     =  (~samples)                  torch.Size([479])
GENERATE:       INFO:     Torch size of rna_labels_new =  (~samples)                  torch.Size([479])
GENERATE:       INFO:     now saving to Torch dictionary (this takes a little time)
GENERATE:       INFO:     finished saving Torch dictionary to data/dlbcl_image/train.pth
TRAINLENEJ:     INFO: 1 about to set up Tensorboard
TRAINLENEJ:     INFO:   Tensorboard has been set up
TRAINLENEJ:     INFO: 2 about to load experiment config
TRAINLENEJ:     INFO:   experiment config has been loaded
TRAINLENEJ:     INFO: 3 about to load models VGG11 and DENSE
LENETIMAGE:     INFO        about to call model for genes net
DENSE:          INFO:                input  dimensions (n_genes)   = 60483
DENSE:          INFO:                hidden layer neurons          = 2000
DENSE:          INFO:                output dimensions (n_classes) = 8
DENSE:          INFO:                dropout (proportion)          = 0.1
```

6 Data Analysis

Performs correlation and clustering analyses on input gene sets, including at present, correlation, covariance, principle component analysis and K-means clustering and provides graphical visualizations of the outcomes. Used in conjunction with Autoencoder dimensionality reduction.

System Hardware

Most development is carried out on a high end ‘MSI GS76 Stealth’ laptop with a 6 core 12 thread Intel 9th Gen CPU, NVIDIA GeForce 2080 RTX GPU; 32GB RAM and a Samsung 1TB (soon 3TB) M.2 solid state drive.

Experiments are conducted on a water cooled tower PC equipped with a 16 core / 32 thread AMD Ryzen Threadripper 3950X; 2 x bridged NVIDIA TITAN RTX GPUs; 128GB RAM; a 2TB M.2solid state drive (‘SSD’) and 4TB of conventional HDD.

Some experiments have also been conducted on Amazon’s AWS cloud. In general this proved not to be as convenient for development and experiment efficiency as local hardware.

Should the local experiment system prove to be inadequate for larger scale experiments, they could alternatively be carried out on PMCC’s GPU cluster, which is CUDA based. At the moment it’s not clear if this will be necessary or not.

Turnitin originality report:

It is strongly recommended that candidates prepare, with their supervisor(s), a Turnitin originality report for a substantial piece of writing (eg. a chapter, a section of a chapter, or the written progress report). To access Turnitin go to the Thesis Similarity Checking community under ‘My Communities’ on the LMS. There is a link to the LMS under ‘Learning Tools’ on the ‘Home’ tab of the Student Portal (http://portal.unimelb.edu.au).

[1] Salto-Tellez and Cree. *Cancer Taxonomy: Pathology Beyond Pathology*. European Journal of Cancer. **115** 57-60 (2019). doi.org/10.1016/j.ejca.2019.03.026

[2] Bright. *Cancer: Its Classification And Remedies*. S W Butler (1871). Google Books.

[3] Burney I. *A Historical Tale of Two Lymphomas: Part Ii: Non-Hodgkin Lymphoma*. Sultan Qaboos University Med Journal. **15**(3) 317-321 (2015) Doi: 10.18295/squmj.2015.15.03.003

[4] Medeiros, Elenitoba-Johnson, *Anaplastic Large Cell Lymphoma.* American Journal of Clinical Pathology, **127** (5) 707–negative722 (2007). DOI: 10.1309/R2Q9CCUVTLRYCF3H

[5] N.L. Harris, E.S. Jaffe, J. Diebold, G. Flandrin, H.K. Muller-Hermelink, J. Vardiman. *Lymphoma Classification – From Controversy to Consensus: the R.E.A.L. and WHO Classification of Lymphoid Neoplasms*. Annals of Oncology, **11**, Supplement 1, S3-S10.(2000) 10.1093/annonc/11.suppl_1.S3

[6] Golub. *Toward a Functional Taxonomy of Cancer*. Cancer Cell **6**, (2) (2004) 107-108. DOI: 10.1016/j.ccr.2004.08.007

[7] Faltas. *Lumpers and Splitters: A New Molecular Taxonomy for Cancer*. Science Translational Medicine. **6,** (249) 249-138 (2014). DOI: 10.1126/scitranslmed.3010118

[8] Hoadley, Yau, Wolf et al. *Multiplatform Analysis of 12 Cancer Types Reveals Molecular Classification within and across Tissues of Origin*. Cell **158** (4) 929-944 (2014). DOI: 10.1016/j.cell.2014.06.049

[9] Tsang and Tse. *Molecular Classification of Breast Cancer.* Advances In Anatomic Pathology, **27**,1, pp. 27-35(9). (2020) DOI: 10.1097/PAP.0000000000000232

[10] Golub, Slonim, Tamayo et al. *Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring*. Science 15 Oct 1999 : 531-537. DOI: 10.1126/science.286.5439.531

[11] Wlodarska, De Wolf-Peeters et al. *The Cryptic inv(2)(p23q35) Defines a New Molecular Genetic Subtype of ALK-Positive Anaplastic Large-Cell Lymphoma*. Blood 1998; **92** (8): 2688–2695 (1998). DOI: 10.1182/blood.V92.8.2688

[12] Jaffe, Barr, Smith. *Understanding the New WHO Classification of Lymphoid Malignancies: Why It's Important and How It Will Affect Practice*. American Society of Clinical Oncolology Educational Book. 37 535-546 (2017) doi:10.1200/EDBK_175437

[13] Scott E. Miller. *DNA Barcoding and the Renaissance of Taxonomy*. Proceedings of the National Academy of Sciences. **104** (12) 4775-4776 (2007) . DOI: 10.1073/pnas.0700466104

[14] LeCun, Y., Bengio, Y. & Hinton, G. *Deep Learning*. *Nature* **521,** 436–444 (2015). DOI: 10.1038/nature14539

[15] Sidike, Alom,Taha, Asari. *The History Began from AlexNet: A Comprehensive Survey on Deep Learning Approaches.* (2018) https://arxiv.org/abs/1803.01164

[16] Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones. *Opportunities and Obstacles For Deep Learning In Biology and Medicine*. Journal of the Royal Society Interface. **15**, 20170387 (2018). DOI: 10.1098/rsif.2017.0387

[17] Geert Litjens, Thijs Kooi, Babak Ehteshami, Bejnordi. *A Survey On Deep Learning In Medical Image Analysis.* Medical Image Analysis **42**, 60-88 (2017). DOI: 10.1016/j.media.2017.07.005

[18] Deng, S., Zhang, X., Yan, W. et al. *Deep Learning in Digital Pathology Image Analysis: A Survey. Frontiers of Medicine*. **14**, 470–487 (2020). 10.1007/s11684-020-0782-9

[19] Serag, Ion-Margineanu, Qureshi, McMillan. *Translational AI and Deep Learning in Diagnostic Pathology.* Frontiers in Medicine (2019). DOI: 10.3389/fmed.2019.00185

[20] Cong L, Feng W, Yao Z, Zhou X, Xiao W. *Deep Learning Model as a New Trend in Computer-aided Diagnosis of Tumor Pathology for Lung Cancer.* J ournal iof Cancer **11**(12):3615-3622.(2020) DOI: 10.7150/jca.43268

[21] Azuaje, F. *Artificial Intelligence For Precision Oncology: Beyond Patient Stratification*. Nature Precision Oncology **3**, 6 (2019). DOI: 10.1038/s41698-019-0078-1

[22] Trister AD. *The Tipping Point for Deep Learning in Oncology*. *JAMA Oncology* **5,** 10 1429–1430 (2019). DOI: 10.1001/jamaoncol.2019.1799

[23] Zaidan, Zaidan, Albahri et al. *A Review On Smartphone Skin Cancer Diagnosis Apps in Evaluation And Benchmarking: Coherent Taxonomy, Open Issues and Recommendation Pathway Solution*. Health Technologies **8**, 223–238 (2018). DOI: 10.1007/s12553-018-0223-9

[24] Esteva, A., Kuprel, B., Novoa, R. et al. *Dermatologist-Level Classification Of Skin Cancer With Deep Neural Networks*. *Nature* **542,** 115–118 (2017). DOI: 10.1038/nature21056

[25] Lo Ying-Chih; Keng-Hung, Lin; Bair, Henry; Sheu Wayne Huey-Herng; Chi-Sen, Chang et al. *Epiretinal Membrane Detection at the Ophthalmologist Level using Deep Learning of Optical Coherence Tomography*. Scientific Reports (Nature Publisher Group); London. **10**, 1 (2020). DOI:10.1038/s41598-020-65405-2

[26] Deng, S., Zhang, X., Yan, W. et al. *Deep Learning in Digital Pathology Image Analysis: A Survey*. Frontiers of Medicine. **14,** 470–487 (2020). DOI: doi.org/10.1007/s11684-020-0782-9

[27] Jiang, Y, Yang, M, Wang, S, Li, X, Sun, Y. *Emerging role of deep learning-based artificial intelligence in tumor pathology*. *Cancer Communications*. **40,** 154– 166 (2020). DOI: 10.1002/cac2.12012

[28] Ping Luo, Yulian Ding, Xiujuan Lei, and Fang-Xiang Wu. *deepDriver: Predicting Cancer Driver Genes Based on Somatic Mutations Using Deep Convolutional Neural Networks*. Frontiers in Genetics. **10**, 13 (2019). DOI: 10.3389/fgene.2019.00013

[29] Kather, J.N., Heij, L.R., Grabsch, H.I. *et al.* *Pan-Cancer Image-Based Detection of Clinically Actionable Genetic Alterations*. *Nat Cancer* **1,** 789–799 (2020). DOI: 10.1038/s43018-020-0087-6

[30] Benoît Schmauch, Alberto Romagnoni, Elodie Pronier et al. *Transcriptomic Learning for Digital Pathology*. BioArxiv (2019). DOI: 10.1101/760173

[31] Schmauch, B., Romagnoni, A., Pronier, E. *et al.* A Deep Learning Model to Predict rna-seq Expression Of Tumours from Whole Slide Images. Nature Communications **11,** 3877 (2020). 10.1038/s41467-020-17678-4

[32] Gundersen, G, Dumitrascu, B, Engelhardt, B. *End-To-End Training of Deep Probabilistic CCA on Paired Biomedical Observations.* In proceedings of the Conference on Uncertainty in Artificial Intelligence. (2019)

[33] Ainscough, B.J., Barnell, E.K., Ronning, P. *et al. A Deep Learning Approach to Automate Refinement Of Somatic Variant Calling from Cancer Sequencing Data*. *Nature Genetics* **50,** 1735–1743 (2018). 10.1038/s41588-018-0257-y

[34] Xiaoli Wang , Jingjing Wu, MiAnika Cheerla, Olivier Gevaert. Deep Learning With Multimodal Representation for Pancancer Prognosis Prediction. Bioinformatics, **35**, 14, i446–i454. DOI: 10.1093/bioinformatics/btz342ngzhi Zhang.

[35] Le H, Gupta R, Hou L, et al. *Utilizing Automated Breast Cancer Detection to Identify Spatial Distributions of Tumor-Infiltrating Lymphocytes in Invasive Breast Cancer*. *American Journal of Pathology* **190**, 7 1491-1504 (2020). 10.1016/j.ajpath.2020.03.012

[36] Shahira Abousamra, Le Hou, Rajarsi Gupta et al, *Learning from Thresholds: Fully Automated Classification of Tumor Infiltrating Lymphocytes for Multiple Cancer Types* (July 2019) https://arxiv.org/abs/1907.03960

[37] Shaver, M.M., Kohanteb, P.A., Chiou, C., Bardis, M.D., Chantaduly, C., Bota, D., Filippi, C.G., Weinberg, B., Grinband, J., Chow, D.S., Chang, P.D. *Optimizing Neuro-Oncology Imaging: A Review of Deep Learning Approaches for Glioma Imaging*. Cancers **11**, 829 (2019) 10.3390/cancers11060829

[38] Aïcha BenTaieb, Ghassan Hamarneh. *Deep Learning Models for Digital Pathology*. Arxiv Computer Vision and Pattern Recognition (2019). https://arxiv.org/abs/1910.12329

[39] Diamantidis, Papadopoulos, Kaiafa, G. et al. *Differential Diagnosis and Treatment of Primary, Cutaneous, Anaplastic Large Cell Lymphoma: Not Always an Easy Task*. Int Journal of Hematology **90**, 226–229 (2009). DOI: 10.1007/s12185-009-0365-7

[40] Erik Peterson, Jason Weed, Kristen Lo Sicco, Jo-Ann Latkowski, *Cutaneous T Cell Lymphoma: A Difficult Diagnosis Demystified*, Dermatologic Clinics, **37**, 4 (2019) 455-469, 10.1016/j.det.2019.05.007

[41] Win, Khin Than1; Liau, Jau-Yu2; Chen, Bo-Jung et al. *Primary Cutaneous Extranodal Natural Killer/T-Cell Lymphoma Misdiagnosed as Peripheral T-Cell Lymphoma: The Importance of Consultation/Referral and Inclusion of EBV In Situ Hybridization for Diagnosis*. Applied Immunohistochemistry & Molecular Morphology. **24**, 2, 105-111(7) (2016) DOI: 10.1097/PAI.0000000000000162

[42] Savopoulos CG, Tsesmeli NE, Kaiafa GD, et al. *Primary Pancreatic Anaplastic Large Cell Lymphoma, ALK Negative: a Case Report*. World Journal of Gastroenterology. **11** (39) 6221-6224. DOI: 10.3748/wjg.v11.i39.6221

[43] Daniel Benharroch, Zarouhie Meguerian-Bedoyan, Laurence Lamant,et al. *ALK-Positive Lymphoma: A Single Disease With a Broad Spectrum of Morphology*. Blood;**91** (6): 2076–2084. 1998 DOI: 10.1182/blood.V91.6.2076

[44] Vassallo, Lamant, Brugieres, et al. *ALK-Positive Anaplastic Large Cell Lymphoma Mimicking Nodular Sclerosis Hodgkin's Lymphoma*. The American Journal of Surgical Pathology **30**, 2, 223-229 (2006) DOI: 10.1097/01.pas.0000179123.66748.c2

[45] Yeo-Rye Cho, Jeong-Wan Seo, Sung Yong Oh, Min-Kyoung Pak, Ki-Ho Kim. *The expressions of MUM-1 and Bcl-6 in ALK-negative systemic anaplastic large cell lymphoma with skin involvement and primary cutaneous anaplastic large cell lymphoma*. International Journal of Clinical and Experimental Pathology. 2020; **13**(7): 1682–1687.

[46] Xiaoli Wang , Jingjing Wu, MiAnika Cheerla, Olivier Gevaert. Deep Learning With Multimodal Representation for Pancancer Prognosis Prediction. Bioinformatics, **35**, 14, i446–i454. DOI: 10.1093/bioinformatics/btz342ngzhi Zhang.

[47] Savage, Harris, Vose et al, for the International Peripheral T-Cell Lymphoma Project, *ALK− Anaplastic Large-Cell Lymphoma Is Clinically and Immunophenotypically Different From Both ALK+ ALCLand Peripheral T-Cell Lymphoma, Not Otherwise Specified: Report From the International Peripheral T-Cell Lymphoma Project*. Blood111 (12): 5496–5504 (2008). DOI: 10.1182/blood-2008-01-134270

[48] Delsol, Brugières, Gaulard et al. Anaplastic Large Cell Lymphoma, ALK-Positive And Anaplastic Large Cell Lymphoma ALK-Negative. Hematology Meeting Reports (formerly Haematologica Reports), **3**(1) (2009) DOI: 10.4081/hmr.v3i1.530

[49] Karen Simonyan, Andrew Zisserman. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. Published as a conference paper at the International Conference on Learning Representations 2015.

[50] Kaiming He, Xiangyu Zhang, Shaoqing Ren, Jian Sun. *Deep Residual Learning for Image Recognition*. Published as a conference paper at the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). DOI: 10.1109/CVPR.2016.90

[51] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, Andrew Rabinovich. *Going Deeper With Convolutions*. Published in 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). DOI: 10.1109/CVPR.2015.7298594

[52] Tensorflow: 'An end-to-end open source machine learning platform' Web site: https://www.tensorflow.org/

[53] Pytorch. 'An open source machine learning framework that accelerates the path from research prototyping to production deployment.' PyTorch project web site: https://pytorch.org/

[54] 'Caffe is a deep learning framework made with expression, speed, and modularity in mind. It is developed by Berkeley AI Research (BAIR) and by community contributors.' Caffe web site: https://caffe.berkeleyvision.org/

[55] The Cancer Genome Atlas (TCGA) Data Portal home page: https://portal.gdc.cancer.gov/

[56] Grand Challenge home page: "A platform for end-to-end development of machine learning solutions in biomedical imaging." https://grand-challenge.org/

[57] CAMELYON17 Grand Challenge home page: https://camelyon17.grand-challenge.org/

[58] Guilherme Aresta, Teresa Araújo, Scotty Kwok, et al. BACH: Grand Challenge on Breast Cancer Histology Images, Medical Image Analysis, **56**, 122-139. (2019) 10.1016/j.media.2019.05.010

[59] Campanella, G., Hanna, M.G., Geneslaw, L. et al. *Clinical-Grade Computational Pathology Using Weakly Supervised Whole Deep Learning On Whole Slide Images*. Nature Medicine **25**, 1301–1309 (2019). DOI: 10.1038/s41591-019-0508-1

[60] Coudray, N., Ocampo, P.S., Sakellaropoulos, T. et al. *Classification And Mutation Prediction From Non–Small Cell Lung Cancer Histopathology Images Using Deep Learning*. Nature Medicine **24**, 1559–1567 (2018). DOI: 10.1038/s41591-018-0177

[61] N. Brancati, G. De Pietro, M. Frucci and D. Riccio, A Deep Learning Approach for Breast Invasive Ductal Carcinoma Detection and Lymphoma Multi-Classification in Histological Images, in IEEE Access, vol. **7,** pp. 44709-44720, 2019. DOI 10.1109/ACCESS.2019.2908724.

[62] Han Le, Rajarsi Gupta, Le Hou, Shahira Abousamra, Danielle Fassler, Tahsin Kurc, Dimitris Samaras, Rebecca Batiste, Tianhao Zhao, Arvind Rao, Alison L. Van Dyke, Ashish Sharma, Erich Bremer, Jonas S.Almeida, Joel Saltz. *Utilizing Automated Breast Cancer Detection to Identify Spatial Distributions of Tumor Infiltrating Lymphocytes in Invasive Breast Cancer*. American Journal of Pathology. 2020 Jul; **190** (7):1491-1504. DOI: 10.1016/j.ajpath.2020.03.012

[63] Liron Pantanowitz, Gabriela M Quiroga-Garza, Lilach Bien et al. *An Artificial Intelligence Algorithm For Prostate Cancer Diagnosis in Whole Slide Images of Core Needle Biopsies: A Blinded Clinical Validation and Deployment Study. **2**, 8 E407-E416 (2020). DOI:* 10.1016/S2589-7500(20)30159-X

[64] Daisuke Komura, Shumpei Ishikawa. *Machine Learning Methods for Histopathological Image Analysis.* Computational and Structural Biotechnology Journal, **16**, 34-42 (2018). DOI: doi.org/10.1016/j.csbj.2018.01.001

[65] Zhiqiang Zhang, Yi Zhao, Xiangke Liao, Wenqiang Shi et al, *deep learning in omics: a survey and guideline*. Briefings in Functional Genomics, **18**, 1, 41–57. (2019) DOI: 10.1093/bfgp/ely030

[66] Feng Gao, Wei Wang , Miaomiao Tan et al. *DeepCC: A Novel Deep Learning-Based Framework For Cancer Molecular Subtype Classification*. Oncogenesis **8**, 44 (2019). DOI: 10.1038/s41389-019-0157-8

[67] Joel S. Parker, Michael Mullins, Maggie C.U. Cheang et al. *Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes*. Journal of Clinical Oncology **27**:8, 1160-1167 (2009). DOI: 10.1200/JCO.2008.18.1370

[68] MsigDB home page. https://www.gsea-msigdb.org/gsea/msigdb/

[69] Sangseon Lee, Sangsoo Lim, Taeheon Lee, Inyoung Sung, Sun Kim. *Cancer Subtype Classification and Modeling By Pathway Attention and Propagation*. Bioinformatics, **36,** Issue 12, 3818–3824 (2020). DOI: 10.1093/bioinformatics/btaa203

[70] Minoru Kanehisa, Susumu Goto. *KEGG: Kyoto Encyclopedia of Genes and Genomes*. Nucleic Acids Research, **28**, 127–30 (2000). DOI: 10.1093/nar/28.1.27

[71] Joshua J. Levy , Alexander J. Titus, Curtis L. Petersen, Youdinghuan Chen, Lucas A. Salas, Brock C. Christensen. *MethylNet: an automated and modular deep learning approach for DNA methylation analysis*. BMC Bioinformatics **21**, 108 (2020). DOI: 10.1186/s12859-020-3443-8

[72] Abdullah Al Mamun, Ananda Mohan Mondal. *Long Non-coding RNA Based Cancer Classification using Deep Neural Networks*. In Proceedings of the 10th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '19) (2019). Association for Computing Machinery. DOI: 10.1145/3307339.3343249

[73] Xena Functional Genomics Explorer home page: https://xenabrowser.net/

[74] Gundersen, G, Dumitrascu, B, Engelhardt, B. *End-To-End Training of Deep Probabilistic CCA on Paired Biomedical Observations*. In proceedings of the Conference on Uncertainty in Artificial Intelligence. (2019)

[75] Gene Tissue Expression (GTEx) Dataset home page: https://gtexportal.org/home/

[76] Carmichael, Calhoun, Hoadley. *Joint And Individual Analysis Of Breast Cancer Histologic Images And Genomic Covariates* (2019).

[77] Qing Feng, Meilei Jiang, Jan Hannig, J.S. Marron. *Angle-Based Joint and Individual Variation Explained*. Journal of Multivariate Analysis, **166**, 241-265 (2018). DOI: 10.1016/j.jmva.2018.03.008

[78] Ash, Darnell, Munro, Engelhardt. *Joint Analysis of Gene Expression Levels and Histological Images Identifies Genes Associated with Tissue Morphology*. BioArxiv pre-print (2018). DOI: doi.org/10.1101/458711

[79] Anika Cheerla, Olivier Gevaert. Deep Learning With Multimodal Representation for Pancancer Prognosis Prediction. Bioinformatics, **35**, 14, i446–i454. DOI: 10.1093/bioinformatics/btz342

[80] Islam, Huang, Ajwad et al, An Integrative Deep Learning Framework for Classifying Molecular Subtypes of Breast Cancer, Computational & Structural Biotechnology Journal, **18**, 2185-2199, (2020). DOI: 10.1016/j.csbj.2020.08.005

[81] Campbell, P.J., Getz, G., Korbel, J.O. et al. *Pan-Cancer Analysis of Whole Genomes*. Nature **578**, 82–93 (2020). 10.1038/s41586-020-1969-6