A MULTI-TASK MACHINE LEARNING PIPELINE FOR THE CLASSIFICATION AND ANALYSIS OF CANCERS FROM GENE EXPRESSION DATA

by

Michael Disyak

B.Sc., Trent University, 2016

B.Sc., Brock University, 2011

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

MASTER OF SCIENCE

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES

(Bioinformatics)

THE UNIVERSITY OF BRITISH COLUMBIA

(Vancouver)

February 2021

The following individuals certify that they have read, and recommend to the Faculty of Graduate and Postdoctoral Studies for acceptance, a thesis entitled:

A Multi-task Machine Learning Pipeline for the Classification and Analysis of Cancers from Gene Expression Data

submitted by **Michael Disyak** in partial fulfillment of the requirements for the degree of **Master of Science** in **Bioinformatics**.

Examining Committee:

Dr. Steven Jones, Professor, Department of Medical Genetics, UBC

Supervisor

Dr. Inanc Birol, Professor, Department of Medical Genetics, UBC

Supervisory Committee Member

Dr. Andrew Roth, Assistant Professor, Department of Molecular Oncology, UBC

Supervisory Committee Member

Abstract

The work contained within this thesis sought to accurately classify 55 primary cancer subtypes, 20 metastatic cancer subtypes, and 16 normal tissues using gene expression data. The classification was done using a multiple learning task approach in which an artificial neural network model makes four distinct classifications at varying levels of biological hierarchy for each input sample. These learning tasks were the organ system of origin, the disease state, the cancer type, and the cancer subtype. The model achieved classification performance ranging from a macro F1-score of 0.987 within the disease state learning task to 0.831 within the cancer subtype learning task on a test set composed of primary cancer, metastatic cancer, and normal tissue samples.

Having shown good classification performance of the model, the second part of the thesis focused on leveraging what the model has learned to extract biological information about the various cancers present in the data set. A backpropagation-based tool called DeepLift was used to generate a list of importance scores for each gene within every class of each learning task. The list of scores was then analyzed for trends that could be utilized to infer biological insight about specific cancer types and subtypes, and between primary and metastatic cancers as individual groups. The lists provide a means to functionally annotate enriched pathways and to quantify and compare the role of RNA genes and pseudogenes across various classes and learning tasks. Some of the results output by DeepLift were validated for their biological relevance by presenting supporting evidence from relevant scientific literature. The ultimate product of this thesis research is a tool with which one can quantify the role of a

variety of genes within cancers spanning both primary and metastatic cancer types. Further analysis of the output generated by the tool could provide a better understanding of the role of genetic expression, including RNA and pseudogenes, within a variety of different cancers.

Lay Summary

The purpose of this thesis work was to leverage machine learning to learn about a variety of cancers from their gene expression data. A machine learning model was created that was able to accurately classify a variety of cancers. Once the model was validated for sufficient accuracy and performance, a second tool was utilized to determine the importance of every gene used by the model in determining the classification for each type of cancer. By examining which genes were indicated as important and their relative rankings, insight into the role of different types of genes and their functions in cancer was investigated. The significance of the genes identified was supported by relevant scientific literature. The combination of tools utilized in this thesis and the output it produces was established as a source of data with which we can improve our understanding of cancer biology.

Preface

This thesis work was conducted under the supervision of Dr. Steven Jones at Canada's Michael Smith Genome Sciences Centre. No explicit ethics approval was required or received for this thesis work. However, this work utilizes data from the Personalized OncoGenomics (POG) project which was approved by and conducted under the University of British Columbia – British Columbia Cancer Agency Research Ethics Board (H12-00137, H14-00681), and approved by the institutional review board (IRB). The POG program whose data is used herein is registered under clinical trial number NCT02155621. Patients involved in the POG program have given consent for tumour profiling using RNASeq as well as whole-genome sequencing.

The thesis approach was designed by me with inspiration for the idea coming from Jasleen Grewal. I conducted all of the experiments contained herein myself. All of the external code libraries used to generated this thesis work and all of the data sources have been referenced appropriately. Where no reference is given, the work is all my own.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

ANN - artificial neural network

BC - British Columbia

CAGE - cap analysis of gene expression

cDNA - complementary deoxyribonucleic acid

DNA - deoxyribonucleic acid

EMT - epithelial–mesenchymal transition

FPKM - fragments per kilobase million

GPH - Genomics Precision Health Database

GSC - Canada's Michael Smith Genome Sciences Centre at BC Cancer

GTEx - Genotype-Tissue Expression Project

IG - immunoglobulin

KEGG - Kyoto Encyclopedia of Genes and Genomes

lncRNA - long non-coding RNA

miRNA - microRNA

ML - machine learning

MLP - multilayer perceptron

mRNA - messenger ribonucleic acid

MRP - mitoribosomal proteins

NCI - National Cancer Institute

ncRNA - non-coding RNA

NGS - next-generation sequencing

NIH - National Institute of Health

NN - neural network

POG - Personalized OncoGenomics (POG) program at BC Cancer

RNA - ribonucleic acid

RNA-seq - ribonucleic acid sequencing

RPKM - reads per kilobase million

SAGE - serial analysis of gene expression

SGD - stochastic gradient descent

TARGET - Therapeutically Applicable Research to Generate Effective Treatments project

TCGA - The Cancer Genome Atlas

TFRI - Terry Fox Research Institute

TPM - transcripts per kilobase million

Acknowledgements

I want to acknowledge all of the great people at Canada's Michael Smith Genome Sciences Centre, with a particular emphasis on my supervisor Dr. Steven Jones and my fellow lab mates. Thank you for the support and answering all of my random questions. Without you all I would not have been able to complete this thesis. Karen Mungall deserves a shout-out here as well. Without her support, I would not have entered into the Masters program at UBC, and I would not have met all the great people at the GSC. Thank you Karen!

I also want to give special acknowledgement to Jasleen Grewal for helping me get this work off the ground. I hope you achieve all of your ambitions. You have certainly helped me reach mine. Thank you to Jean-Michel Garant for helping me with my RNA understanding and analysis. I wish you the best of luck with your professional pursuits (and all those cats). Thank you to Luka Culibrk for your helpful questions. You sir, are a genius. Thank you to Emre Erhan for being so nice and connecting me with so many people, keep on crushing it. Finally, thank you to Jenny Yang for being my lab bestie. Good luck on your PhD journey!

Of course, my thanks also extends to my whole family but especially to Pepa, Pop, Sascha, and Opa. Thank you all for the support while I finally complete my school journey. Love you guys.

Last but not least, thank you to my supervisory committee for taking the time to help me achieve my goals. Your time and effort is much appreciated.

Chapter 1

Introduction

The purpose of this thesis is to attempt to create a machine learning tool that can aid in the diagnosis of cancers from gene expression (transcriptome) data and subsequently leverage this tool to better characterize the underlying biology. The approach taken here leverages machine learning in a multiple learning task approach. The four learning tasks selected for the machine learning model represent a biological hierarchy that may help the model to better classify cancers. If a model can be trained to understand the features of cancers at the gene expression level, we can then work to extract any insights the model has gleaned. Ultimately, the goal of this work is to characterise and quantify (where possible) the role of gene expression in a variety of cancers.

The focus of this research can be summarized by the three goals below:

1. Create a multi-task neural network model to accurately classify four categories of biological interest from gene expression data:
    - Organ System of Origin
    - Disease State: primary cancer, metastatic cancer, or normal tissue
    - Cancer Type

- Cancer Subtype

2. Identify and extract the genes utilized by the model to determine the classification of each category

3. Utilize the identified genes to validate and infer biological information about cancer

The following sections will introduce some important background information to motivate this thesis work and place it in the current context of cancer research and machine learning.

1.1 Background of Cancer

Cancer is a group of diseases defined by the abnormal growth of cells [**1**]. This uncontrolled growth is often caused by acquired or inherent (somatic) genetic mutations that circumvent the normal cell life cycle resulting in the formation of abnormal tissue growth (tumours). The abnormal growth is driven by mutations that inhibit cell growth suppressors, activate growth factors, and/or improve cell proliferation and motility [**1**]. Identifying mutations responsible for driving tumourigenesis (the creation of tumours) is the focus of many research endeavours world-wide, including this thesis.

Cancer is the second leading cause of death in Canada and the world [**1**, **2**]. In 2018, it accounted for one sixth of all mortalities world-wide (9.5 million deaths) and there are 83,300 cancer-related deaths expected in Canada in 2020 [**1**, **2**]. It is estimated that almost half of Canadians will be diagnosed with cancer at some point in their lives, and while the cancer mortality rates have decreased over the last 40 years, the overall number of new cases and fatalities has been increasing along with the average age of Canadians [**2**, **3**]. Clearly, cancer remains a prominent health issue both in Canada and around the world. On-going cancer research should remain a prime focus for improving the health and life-span of human beings.

Cancer-related death is often the result of complications caused by metastasis. In fact, 67-90% of all cancer-related deaths are attributed to metastases (secondary tumours) which are defined by the spread of tumour cells beyond the primary site of origin into the surrounding tissues and/or to distant regions of the body [**1**, **4**, **5**]. Once a tumour has spread to a critical organ, like the brain or lungs, if the growth is not stifled it ultimately results in organ failure and death. As a result of the prominence metastasis plays in cancer deaths, we must prevent the formation and proliferation of metastatic cancers in order to reduce the impact of this disease.

For metastases to arise, the cells from the primary tumour must not only physically disseminate from the primary site, but must adapt to the new micro-environment present at the secondary site [**6**, **7**]. The ability to disseminate and adapt is a key feature of metastatic cancers. It can be postulated that there are genetic characteristics underlying these abilities. In order to properly identify the origin of these abilities and subsequently hinder them, we must be able to effectively characterize the underlying genetic causes of cancers [**6**]. Often there are multiple genetic factors influencing the ability of tumours to spread, and the primary site of origin can predispose some tumours to higher aggression and adaptability [**6**]. For example, tumours of the lungs often spread widely and rapidly, whereas tumours of the prostate and breast are typically much more docile and limited in secondary site proliferation [**8**, **9**]. It is partly for this reason that identifying the site of origin is a key step in diagnosing cancer type and ultimately deciding on the most effective treatment protocol [**6**, **7**].

Prior to the advent of genetic sequencing, we had no ability to directly characterize the genetics of cancer and thus relied solely on morphological and immunohistochemical analysis to determine a cancer's site of origin and type. This approach is problematic as the accuracy of diagnosis using these techniques can be less than ideal , particularly with metastatic cancers [**10**]. A meta-analysis by Anderson and Weiss in 2010 found that only 65% of metastatic cancers had their site of origin correctly identified through immunohistochemical analysis compared to 82% with a mixture of primary and metastatic samples [**10**]. This is

clear evidence for the need to improve our capacity to characterize cancers in new ways. In the era of genetic sequencing, we have the ability to look at the genome of different cancers and attempt to categorize and quantify the role of genetics in the cause and characteristics of various cancers. By leveraging machine learning tools, as exemplified in this thesis, we can aim to identify key genetic markers of cancers and ultimately work to improve diagnosis and treatment.

1.1.1 The Role of Gene Expression in Cancer

The human genome contains two major genomic regions: coding and non-coding regions [**11**]. Coding regions are areas of the genome comprised of genes that encode the information necessary to build proteins from nucleic acids. The level at which protein coding genes are transcribed into RNA is said to be the expression levels of that gene. The set of RNA transcripts generated by both coding and non-coding regions is collectively referred to as the transcriptome [**12**]. The expression levels of genes are a quantitative measure of the rate of transcription of each gene. The rate of transcription can have an impact on the amount of proteins generated from the RNA produced by transcription. Proteins are pivotal to life and the overabundance or unplanned absence of them can cause a myriad of problems including the formation of cancers [**13**]. For this reason, quantifying and analyzing the expression levels of genes is a valuable resource in the cancer research space.

Numerous studies report differential expression of genes as being a potential source of tumourigenesis [**14**, **15**, **16**, **17**]. Over and underexpression of genes, particularly those with functionality linked to cell division, propagation, and apoptosis, are hallmarks of many cancers [**16**, **18**, **19**, **20**]. This knowledge can be leveraged to detect susceptibility to and the characteristics of cancers [**16**, **17**, **21**]. Through categorization of gene expression patterns in cancer types, we can begin to work towards treating the causes and/or effects of differential expression. This thesis work is in part motivated by this goal. If we can detect novel patterns of genetic expression within cancers, we can provide more potential therapeutic targets.

1.2 Background of Genetic Sequencing

Genetic sequencing (DNA sequencing) refers to the determination of the sequence of nucleic acids within a given piece of DNA. The ultimate goal of sequencing is to rapidly and accurately determine the entire sequence of an organism's genome with the intention being to understand the location, composition, and function of all of its genomic regions.

The current state of genetic sequencing arrived as a result of two major breakthroughs. The first of which was the invention of the first-generation of sequencing technology called Sanger sequencing [**22**]. Sanger sequencing was created in 1977 and utilizes radio or dye-labelled chain terminating nucleotides in conjunction with DNA polymerase to grow fragments of the DNA of interest that incorporate the labelled nucleotides [**22**]. By capturing a large enough set of labelled fragments, we will eventually have one fragment with a labelled nucleotide at each position in the DNA sequence of interest. We can then determine which nucleotide exists at each location across all of the fragments and combine the information to obtain the whole DNA sequence. In Sanger sequencing, the visualization process involves gel electrophoresis and is limited by the number of lanes within the gel. It can only sequence one fragment of DNA per gel and can only grow as many dyed fragments as there are lanes in the gel. Furthermore, the labelled nucleotides used are chain-terminating nucleotides which prevents the addition of nucleotides following the dye. Therefore, only a single position in the DNA fragment will be labelled and read. This results in accurate but slow and costly sequencing, particularly when concerning the sequencing of multiple DNA fragments. If one desires to sequence multiple fragments, a gel must be prepared and run for each fragment and a dyed fragment must be produced for each position in the DNA of interest.

The second generation of sequencing technology, also known as next-generation sequencing (NGS), was created in 2006 by Solexa [**22**]. One of the most common forms of NGS today is Illumina sequencing and is the source of all sequence data for this thesis. Illumina sequencing addresses the limitations of Sanger sequencing by allowing multiple reversible dye-labelled

nucleotides to be attached to a single fragment and by not relying on gel electrophoresis to read the fragments. Instead, Illumina sequencing grows millions of DNA fragments simultaneously, each with many dyed nucleotides. This allows for parallel sequencing to be conducted. To accomplish this, Illumina sequencing utilizes a flow cell with millions of wells that can each read a dyed fragment of DNA. The features of Illumina sequencing provide the ability to rapidly sequence an entire genome with a single prepared sample of DNA. This significantly reduces the cost of sequencing both in preparation time and dollars per DNA fragment. For these reasons, it is widely used in genomic studies of cancer.

1.2.1 RNA Sequencing

RNA sequencing (RNA-Seq) refers to determining the presence and order of nucleotides found in an RNA molecule [**24**]. With each iteration of DNA sequencing technology, there have been techniques developed to apply them to RNA as well. The earlier techniques such as serial analysis of gene expression (SAGE) and cap analysis of gene expression (CAGE) shared the limitations present in Sanger sequencing, namely high cost and low throughput [**24**]. Likewise, these limitations have been mitigated significantly with the advent of the second generation of sequencing technology. In order to perform RNA sequencing using Illumina sequencing, the RNA sample goes through an additional sample preparation phase to transcribe the RNA to cDNA (complimentary DNA) using a reverse transcriptase enzyme [**25**]. Having been converted to cDNA, the sample can now undergo the normal Illumina sequencing process. The reads obtained from RNA sequencing are then mapped to a reference genome/transcriptome using one of a number of genome alignment tools such as STAR [**26**]. The number of reads found at each region of the genome are quantified and normalized to determine the expression of that region of the genome.

The purpose of normalization is to overcome potential biases introduced by differing read depth and gene lengths. Normalization to a standard format allows expression data to be compared between studies and attempts to remove technical bias introduced by the sequenc-

ing process [**27**, **28**]. The three most popular normalization formats are: RPKM, FPKM and TPM [**27**]. Each have slightly different ways of implementing normalization. The RPKM format, however, is the relevant format for this thesis work as all of the data used herein was presented as RPKM values. The RPKM value is a within-sample normalization using the reads per kilobase per million reads mapped. It is calculated by dividing the number of reads mapped to each gene by the total number of mapped reads multiplied by the gene length [**27**].

1.3 The Role of RNA Genes in Cancer

A large fraction of the human genome is composed of non-coding regions and have historically been considered "junk" [**29**, **30**]. This non-coding region comprises DNA that is not transcribed to RNA or its RNA transcripts do not code for proteins (non-coding RNAs). In recent years, studies have begun to show the important role non-coding RNAs (ncRNA) play in tumourigenesis and the malignancy of cancers [**29**, **31**, **32**, **33**]. Non-coding RNAs influence many cellular processes involved in cancer, such as cell growth, differentiation, proliferation, and apoptosis [**31**, **32**]. MicroRNAs (miRNA), a class of non-coding RNA, play a key role in these cellular processes by binding messenger RNA (mRNA) targets and either inhibiting or degrading their function [**34**, **35**]. Through this functional modulation, miRNAs are able to affect large changes in gene expression. In fact, miRNAs control nearly one third of all human genes, and for this reason, play an important part in our growing understanding of cancer [**33**]. It should be noted that in the context of cancer, miRNAs are generally considered to be either tumour suppressors or oncomiRs depending on their target mRNAs [**36**]. OncomiRs downregulate tumour suppressor genes and thus act to promote tumourigenesis. In contrast, tumour suppressing miRNAs act to suppress the effects of genes that promote tumours [**36**].

Studies have exemplified the role of miRNAs in cancer and have shown that characterization of cancers is feasible using miRNA biomarkers [**36**, **37**, **38**]. For example, Calin et al

(2002) showed that two miRNA genes (*miR15* and *miR16*) were deleted or downregulated in more than half of their samples of B-cell chronic lymphocytic leukemias (CLL), and thus play a significant role in the pathogenesis of CLL [**37**]. Furthermore, it has been shown that alterations in the miRNA signature between normal and malignant cells can be utilized to accurately classify cancer types and the organ system of origin in poorly differentiated cancers [**33**].

Given the key role RNA genes play in cancer, it is reasonable to expect the model utilized in this thesis to highlight their role. Since the multi-task model used in this work is attempting to learn patterns of expression in cancer, we would expect that some of the genes highlighted by the model should support the findings of previous work on RNA genes in cancer.

1.4 The Role of Pseudogenes in Cancer

Pseudogenes are decayed versions of functional genes that may originate as a result of gene duplication events such as point mutations, insertions, deletions, and/or frameshifts, among others [**39**]. They have, until recently, been considered entirely non-functional regions of the genome and looked upon as "junk DNA" [**40**, **41**]. In recent years, however, studies have shown that thousands of pseudogenes are transcribed and hundreds are translated [**41**]. For this reason, they are considered part of the set of RNA genes known as long non-coding RNA genes (lncRNA). Transcribed pseudogenes can be detected through RNA-seq and have been shown to have diagnostic power as biomarkers in some human cancers [**41**]. A pan-cancer analysis of RNA-seq data has demonstrated that pseudogenes show cancer subtype-specific expression patterns, and in some cases can be used to differentiate between subtypes [**41**]. In light of recent evidence surrounding the utility of pseudogenes in cancer, the implication of pseudogene expression within the context of this thesis will be explored.

1.5 Machine Learning

Machine learning refers to a set of computer algorithms that can independently acquire knowledge and extract patterns from raw data [**22**]. These algorithms are often used to create predictive models. The process of teaching a model to make accurate predictions is referred to as training and requires the use of training data. Training data is a set of data that is representative of the kind of data we wish to understand and make predictions about. Once the model has sufficiently learned a representation of the training data, the model is said to be trained. It can then be used to make predictions about new, unseen data of the same format as the training data. Machine learning models can be particularly useful for pattern recognition tasks in which the data of interest is too large or complex to be sufficiently analyzed by human beings. One such example is the characterization of cancers based on gene expression data, which is the focus of this thesis. There are three major types of machine learning: supervised, unsupervised, and reinforcement learning [**43**]. Supervised learning is the relevant form of machine learning for this thesis work and is described in the following section.

1.5.1 Supervised Learning

Supervised learning is a form of machine learning that utilizes labelled data [**42**, **43**]. This is in contrast to unsupervised learning in which there are no labels present. Labelled data refers to a set of data that contains not only features, but also a label (target) [**42**]. For example, images that are labelled with the type of object being visualized within an image. When the labels are discrete, they are often referred to as classes, and using machine learning to predict classes of data can be referred to as classification. The goal of classification is to generate a model that can accurately label the given training data and make predictions of the labels for new, unseen data [**42**, **43**]. In order for a machine learning model to learn to accurately label the given data, it needs to learn (via the training process) the underlying features that best represent the data for every possible label.

Training a supervised learning model is an iterative process of assigning predictions for each sample of data given in the training set, calculating the loss (error) of these predictions, and adjusting the parameters of the model to reduce the loss, thereby improving accuracy. This process is repeated until the loss reaches a plateau or a desired value. Once a plateau is reached, the model is considered trained and should offer some capacity to accurately predict labels for data previously unseen by the model. The calculation of the loss of a model is discussed in more detail as part of the neural network section below.

In order for a model to accurately classify data, it must learn to accurately represent the data. This requires building a function that takes the input features of the data and transforms it to an output classification. This is the core of any machine learning model. There are many algorithmic processes to produce a representative function and an artificial neural network (ANN or NN) is one example of these. Neural networks are described in the following section and is the relevant class of machine learning models for this thesis.

1.5.2 Neural Networks

A neural network model is at its core a mathematical function. It takes inputs and maps them to outputs. It is composed of a collection of neurons (described in further detail below), often referred to as nodes. These collections of nodes are connected by weighted edges. These weights (along with a bias term for each neuron) are the adjustable parameters learned as part of the training process [**43**]. The process of training a neural network involves iterative updates to the model parameters (weights and biases) in order to reduce the overall error rate (loss) of the model's predictions. This process ultimately results in the model being able to more accurately approximate a representative function of the training data. As the representation improves, the loss of the model should decrease and the prediction accuracy should increase.

The way in which nodes are connected together defines a neural network's architecture and can have a significant effect on its performance [**42**, **44**]. The canonical example of a neural network is a feed-forward neural network where the connections within it are all weighted, directed and acyclic (Figure 1.1) [**42**, **44**]. There are three standard parts to a basic feed-forward network: the input layer, the hidden layer, and the output layer. A layer is simply any number of nodes that exist at the same depth within the network. The depth of a network is the number of layers it contains. The term "deep" is used to refer to networks that have multiple hidden layers [**42**]. The model developed in this thesis contains four hidden layers and is thus considered a deep neural network.

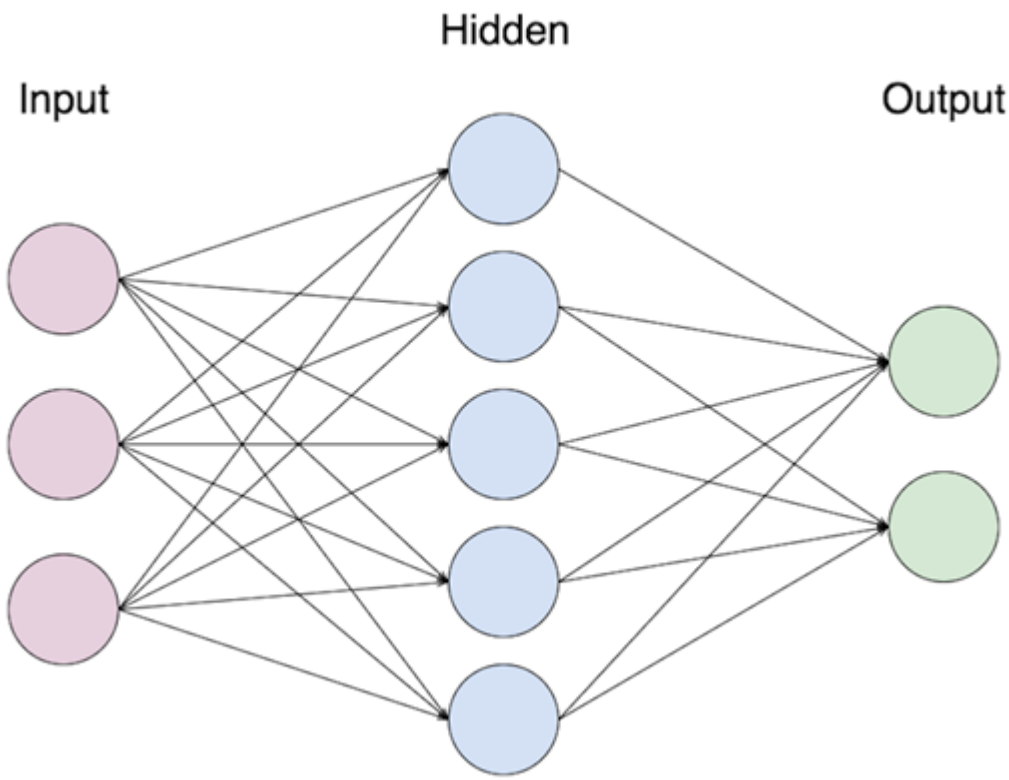

Figure 1.1: A diagram depicting the basic structure and layering of a feed-forward neural network model. This figure was taken from the web [**45**].

The input layer of a neural network is where the data is given to the model and will contain as many nodes as there are input features. In this thesis work for example, the number of input nodes is equal to the number of genes in each sample. In an image classification task,

the input layer would have as many nodes as there are pixels in the image. The hidden layer is named as such because it does not provide information in the form of the desired output [**42**]. The function of the hidden layer is to provide a means with which apply nonlinear transformations to the input. A neural network with a single hidden layer that contains a sufficient number of nodes (width) can approximate any mathematical function [**42**, **44**]. The output layer is the layer of the network where predictions are given. The width of the output layer in a classification task will be such that it can represent the desired number of labels or classes to be assigned to the input data. For example, if a neural network is trying to decide if an image is either a car, a bus, or a truck, the output layer may have 3 nodes. One for a car, one for a bus, and one for a truck. A classification is determined by which corresponding output node has the highest output value (activation). The class being predicted by the model is the one in which the corresponding output node has the highest activation.

1.5.3 Gradient Descent

There are two types of components (parameters) being learned by a neural network during training. These components are the weights w and biases b for each neuron (see Section 1.5.5) within a neural network [**42**, **43**]. One way that these parameters can be learned in a feed-forward neural network is through gradient descent. Gradient descent is the process of calculating the loss of a model followed by updating its parameters in the direction of the negative gradient of that loss. In other-words, it is a descent in the loss of the model through iterative steps in the direction of the negative gradient. The actual calculation of the gradient is done using an algorithm called backpropagation (see the relevant section below) [**42**]. The negative gradient for a neural network is composed of the set of partial derivatives for each parameter and represents the direction of steepest change in the parameters required to minimize the loss of the network. By determining the negative gradient of the model, we can effectively update each parameter to move in the direction that minimizes the loss of the model.

There are two basic forms of gradient descent [**42**, **43**]. Standard batch gradient descent calculates the loss of the model using an average of the losses for every sample (batch) in the training set. Stochastic gradient descent (SGD) uses a single sample selected at random to determine the loss [**43**, **46**]. Stochastic gradient descent can also be performed on multiple samples (mini-batch) whose losses are averaged together [**46**]. This method is referred to as mini-batch gradient descent. Regardless of the number of samples used, the negative gradient is calculated using backpropagation and then used to update each parameter of the model so as to minimize its loss. The formula for updating a parameter can be seen in Equation 1.1 [**42**]. In this equation, x is the current value for a given parameter, x' is the updated value for the same parameter, α is the learning rate, and $\nabla f(x)$ represents the gradient of parameter x.

$$x' = x - \alpha \nabla_x f(x) \tag{1.1}$$

1.5.4 Learning Rate

The learning rate used when training a neural network can have a significant effect on the performance of the model and is often considered the most important hyperparameter [**42**]. The learning rate effects how large of a step in the direction of the negative gradient the model makes for each parameter (see Equation 1.1). The direction and magnitude of the update is determined by the gradient as calculated via backpropagation. The learning rate parameterizes this gradient and can be used to decrease or increase the size of the parameter update. If the rate is too high, the model may not converge to the best solution (minimal loss), and the loss may increase as a result of the update. If we consider the U-shaped generalization error curve (Figure 1.2), we can think of this as overshooting the goal (optimal capacity) [**42**]. If the rate is too small, the model could take a very long time to converge or may not converge to a good solution at all [**42**]. Finding a good learning rate is about balancing the training time with trying to converge as close as possible to the optimal solution for the given architecture and problem space. One technique to balance

these two needs of learning is to use an adaptive learning rate in which training begins with an initial learning rate and is subsequently modified by some algorithmic approach. The adaptation algorithm, the corresponding initial learning rate, the rate/type of reduction, and the frequency at which the reduction rate is applied, are all aspects of the learning rate that must be explored as part of hyperparameter tuning.

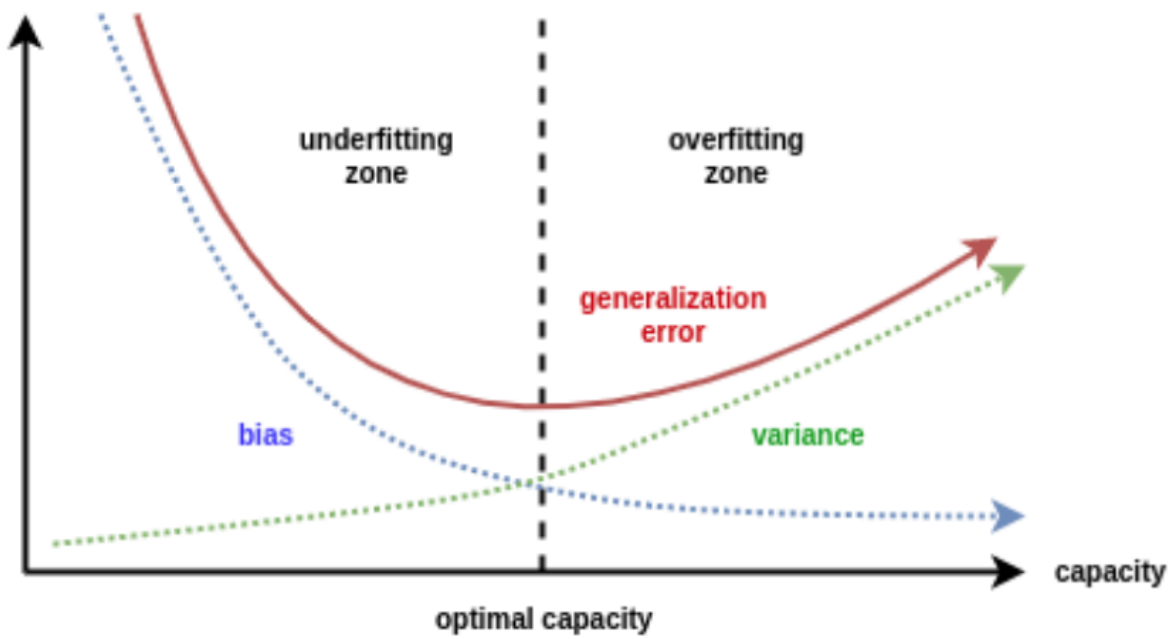

Figure 1.2: A plot depicting the bias-variance trade-off, the U-shaped generalization error curve, the optimal capacity, under-fitting, and over-fitting zones. Note: This figure was taken from Goodfellow et al. (2016) [**42**].

1.5.5 Neuron

Each neuron within a neural network is at its core a mathematical function. To be more specific, it is a nonlinear function [**43**]. The function that comprises a neuron is defined in Equation 1.2 [**43**]. In this equation, x is the input feature vector (all the incoming connections to the neuron), w are the weights associated with each input feature connection, b is a bias term, and σ is an activation function applied to the output of the linear equation $(w \times x + b)$. Also note the summation over each $w \times x$ which makes this term a weighted sum of the input

features. The activation function σ of this equation is used to transform the output of the contained linear combination of inputs $(w \times x + b)$ into either a value between 0 and 1 or a value between -1 and 1 depending on the activation function used [**42**, **43**]. The resulting value $f(x)$ for a neuron is considered its "activation".

$$f(x) = \sigma(\sum(w \times x) + b) \tag{1.2}$$

1.5.6 Activation Functions

Activation functions within the nodes of a neural network serve to apply a non-linear transformation to the output of the linear combination within a neuron (see Section 1.5.5) [**42**]. They are selected on the basis of the task at hand [**43**]. The two relevant activation functions for this thesis are the hyperbolic tangent and softmax functions. These are described below.

Hyperbolic Tangent Function

The Hyperbolic Tangent (tanh) function is a mathematical function suitable for use as an activation function with the nodes of a neural network. It has properties such that $-1 < tanh(x) < 1$ and $tanh(0) = 0$ that make it ideal for use as an activation function within the hidden layers of a neural network [**47**, **43**]. The tanh formula is given below (Equation 1.3) and a graphical representation of it can be seen in figure 1.3 [**48**].

$$tanh(x) = \frac{(e^x - e^{-x})}{(e^x + e^{-x})} \tag{1.3}$$

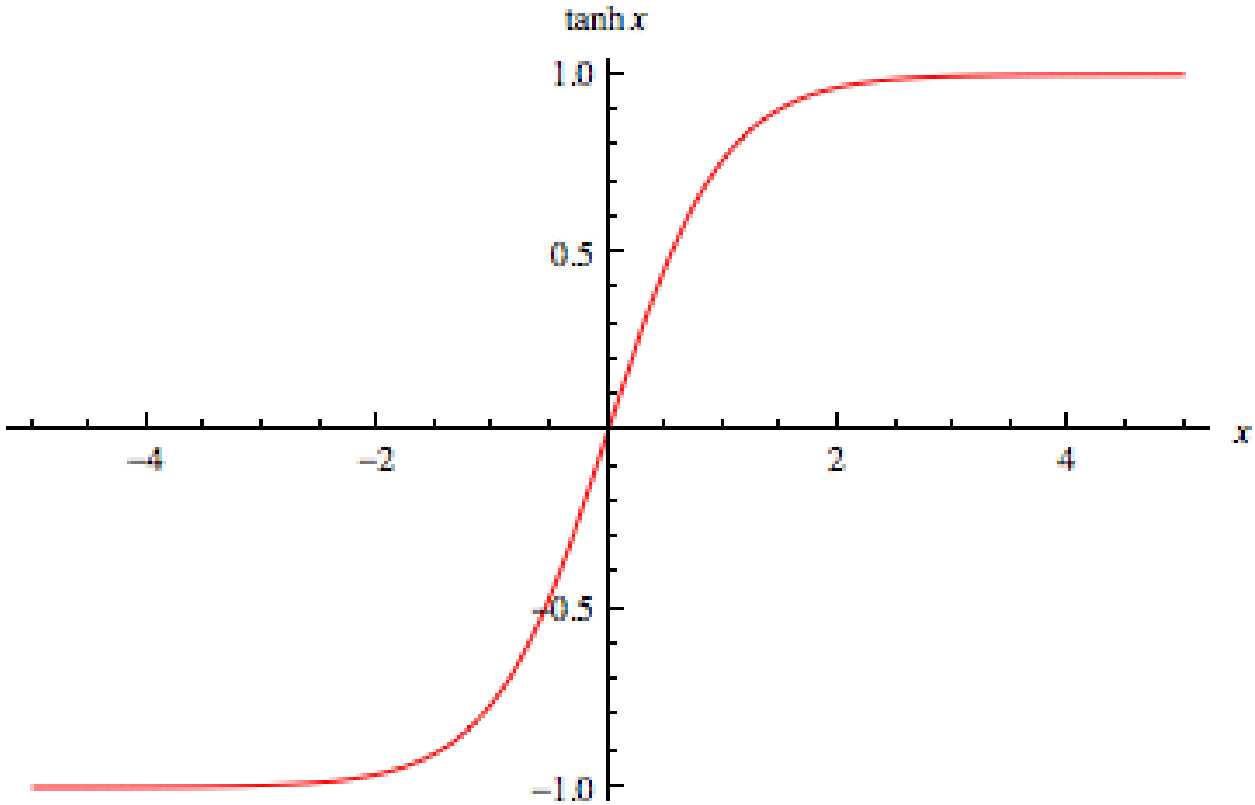

Figure 1.3: A plot of the hyperbolic tangent function. Note: This figure was taken from MathWorld [**48**].

Softmax Function

The softmax function is often used in neural networks within the output layer [**42**]. The softmax function effectively converts the output of the output nodes into a normalized probability-like distribution for each class in the output [**42**]. The output value of each node then corresponds to a percentage of the final classification. For example, if a three-class model using the softmax function in the output layer has output values of 0.2, 0.2, and 0.6, the model is making a 60% classification as the third class and a 20% classification as the first two classes. We would interpret this output as a predicted class of the third type because the third class has the largest output value. The softmax function (σ) is defined in Equation 1.4 below [**42**, **43**]. In this equation, z is the score output for each output node and K is the number of possible classes.

$$\sigma(z)_i = \frac{e^{z_i}}{\sum_{j=1}^{K} e^{z_j}} \text{ for } i = 1, ..., K \text{ and } z = (z_1, ..., z_K) \text{ in } \mathbb{R}^K \tag{1.4}$$

1.5.7 Loss Functions

The loss function (cost function) of a model is used to determine its performance relative to the training data. The loss is a value representative of how much error there is between the predicted and true output of the model [**49**]. The choice of loss function will be dictated by the type of learning task [**42**]. For this thesis work, since a multi-class classification is being conducted, the categorical cross-entropy loss is used. This function is detailed below.

Categorical Cross-Entropy Loss

The categorical cross-entropy (CCE) loss function is the standard loss function used for multi-class classification [**43**, **50**]. The formula for the categorical cross entropy (CE) loss is shown below in Equation 1.5 [**43**, **50**]. In this formula, K is the total number of classes, N is the number of samples, t is the target, and y is the predicted class.

$$CCE = -\sum_{n}^{N} \sum_{c}^{K} t_{nk} \ln y_{nk} \tag{1.5}$$

1.5.8 Backpropagation

Backpropagation refers to the algorithm used for efficiently calculating the gradient of the loss function with respect to each parameter within a neural network [**42**]. Since a neural network is a function composed of other functions (each node in the network is a nonlinear function, see Section 1.5.5), the backpropagation algorithm applies the chain rule (from calculus) in a specific manner to efficiently compute the partial derivative for each composing function [**42**]. The set of partial derivatives for each parameter results in the gradient of the network

as a whole with respect to its loss. These partial derivatives can then be utilized as part of gradient descent to update each parameter and minimize the loss of the network (see Section 1.5.3 for details). Further details on the use of the chain rule for backpropagation can be found in the relevant section of the Deep Learning textbook by Goodfellow, Bengio, and Courville (2016) [**42**].

1.5.9 Initialization

The weights and biases of a neural network need to be initialized to a set of values prior to training. One method of weight initialization is to use Glorot uniform initialization. The Glorot uniform initializer (Equation 1.6) samples from a uniform distribution between the negative and positive limit seen in Equation 1.7 [**51**, **52**]. The biases of a neural network can be and are often simply set to zeros.

$$\text{glorot uniform initializer} = \text{sample[-glorot limit, glorot limit]} \tag{1.6}$$

$$\text{glorot limit} = \sqrt{\frac{6}{\text{number of input nodes} + \text{number of output nodes}}} \tag{1.7}$$

1.5.10 Over-fitting

Over-fitting is a phenomenon in machine learning where a model will learn to represent the training data with increasing accuracy while the accuracy of the model with predictions made on unseen data, such as validation data, decreases [**42**]. There is a point at which the model fits the training data so well that its ability to generalize to unseen data is hindered (see the over-fitting zone in Figure 1.2). Preventing over-fitting is about striking a balance between accurately learning the training data while maintaining an ability to generalize to unseen data. There are many techniques used to prevent over-fitting and these methods are

generally termed regularization methods. The application of regularization to a model is the prevention of over-fitting by virtue of penalizing increasing model complexity [**42**, **53**]. As model complexity goes up, the ability of the model to fit to the training data increases and at a point, the ability to generalize goes down. Two examples of regularization used in this thesis are early stopping and dropout. These are described in their respective sections below.

1.5.11 Early Stopping and Patience

Early stopping refers to halting the training of a machine learning model when a desired metric, often validation loss or accuracy, reaches a minimum or maximum respectively [**42**]. Early stopping is the most common form of regularization used in deep learning and is utilized to help prevent over-fitting [**42**, **54**]. Patience refers to the number of training epochs that will pass before training is halted and is a hyperparameter of the early stopping process that must be selected.

1.5.12 Dropout

Dropout is an effective and computationally inexpensive technique for the regularization of neural networks [**42**, **55**]. It is implemented by setting the weights of connections between a random subset of nodes to 0 at each training step. A visualization of the effect of dropout can be seen in Figure 1.4 [**54**]. Conceptually, dropout can be thought of as training multiple models within a single network [**42**]. By dropping out different connections, we are essentially creating sub-networks at each iteration and forcing the model to learn solutions that do not rely too heavily on any single connection (or set of connections) and thus are more robust [**42**]. The rate of dropout is typically a value between 0 and 1 that indicates what fraction of the connections are set to 0 (dropped out) at any given training iteration.

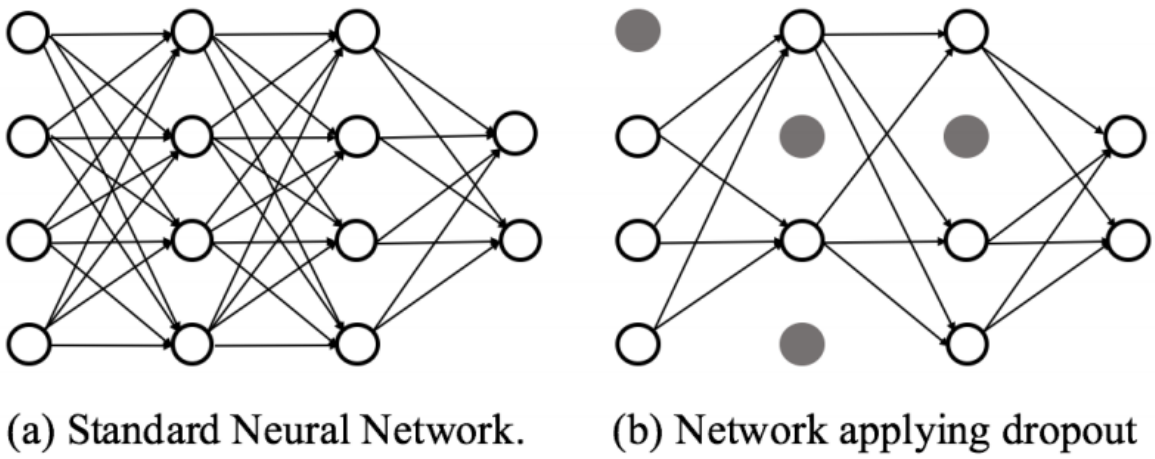

Figure 1.4: (a) A standard fully-connected neural network without dropout. (b) A sub-network created by dropping out some of the connections in the standard neural network. Note: This figure was adapted from Wang et al. (2018) [**54**].

Class Weighting

Class weighting is a method of weighting the training loss that can be used to combat the potentially detrimental effects of training machine learning models using imbalanced data sets [**56**, **57**]. When class weights are appropriately applied, the effect can be to aid in classification performance on minority classes [**58**]. This can be accomplished by weighting the loss for each sample by either an arbitrary value, or by a value that bears some relationship to the size of the class in which the sample belongs. The weight for each class is the class weight and is the value used to weight the loss for each for each sample of that class.

Multi-Task Learning

Multi-task learning refers to machine learning in which there are more than one learning objective being learned in parallel. The effect of multi-task learning is that of improved generalization performance as a result of shared parameters. However, this holds true only

under the assumption that there exists a valid relationship between tasks [**42**]. Each shared parameter is utilized for multiple objectives and therefore the associated parameters are less likely to over-fit the variation in the data related to any singular task [**42**].

1.6 Deeplift

DeepLift is a backpropagation-based tool developed for the interpretation of trained neural networks [**59**]. It uses a backpropagation-like algorithm to determine the effect of a selected set of input nodes on the resulting activation of a set of selected nodes of interest (output nodes in this case). A baseline activation level for the output nodes is established using a reference value for each input node (the default reference is 0). The baseline activation is established by passing the reference value through the network to the output nodes via the selected input nodes. The activation level seen at each output node is then recorded as the baseline activation for that node. The set of training data is then passed through the network inputs. The difference from the reference activation value is calculated at each output node and then propagated back through the network to each input node. The larger the difference from the reference activation caused by a particular input node, the greater the perceived effect of that input node is. The larger the effect of an input node, the higher the score DeepLift will assign to it. Similarly, if an input node reduces the activation of an output node, a negative score is returned. This is repeated for all of the training samples across each input and output node combination. The result is a matrix of positive and negative scores that indicate how important each input node is to the output nodes' activation. In the context of this thesis, we receive a set of scores for each gene in the input data that correspond to how important they are to the classification of each of the output classes. We have essentially asked DeepLift to determine how important each gene is in classifying each of the classes within each learning task of the multi-task model. Further details on how DeepLift works can be found in the DeepLift paper and accompanying videos on YouTube [**59**].

One additional consideration to the analysis of the DeepLift data for this thesis work pertains to gene expression. Since each input feature represents a gene, we must be careful to properly interpret positive scores assigned by DeepLift. A positive importance score for a particular gene does not necessitate that the gene has higher than normal expression. It simply means there is something about the expression of this gene that has a positive influence on the model selecting the current class being examined by DeepLift. This could be under or overexpression of a gene.

Chapter 2

The Data and the Model

2.1 The Data

All of the data used for this thesis consists of RPKM gene expression values for 26668 genes. The genes selected were those that were found in the intersection of all of the genes available across all of the different data sources. The list of these data sources is outlined in Table 2.1. The data can be thought of as two separate sets. The largest set consists of a mix of primary cancer, metastatic cancer, and normal tissue samples. The second, smaller data set contains only metastatic cancer samples and was used only for testing the trained machine learning model.

Primary & Normals
TCGA
NIH-NCI non-Hodgkin lymphoma dataset including FL and DLBC
Non-cell-line GBM data from the TFRI's Glioblastoma Multiforme project
MESO dataset from GenenTech
MB-Adult data from the GSC
Follicular lymphoma data from the GSC
CML data from the TARGET project
CLL and DLBC data from the GPH project
Metastatic
Met500
POG

Table 2.1: List of data sources for primary, normal, and metastatic data

Mixed Data Set

Within the mixed data set, the vast majority of the primary cancer samples are from The Cancer Genome Atlas (TCGA) data set. The TCGA data was supplemented with primary mesothelioma, glioblastoma, non-Hodgkin's lymphoma, medulloblastoma, follicular lymphoma, and leukemia data sets from a variety of other sources detailed in Table 2.1 [**11**]. There are 375 metastatic cancer samples included in the mixed set that came from the Met500 cohort gathered by the University of Michigan. Details of this cohort can be found in the associated paper by Robinson et al. [**62**]. With all of the sources combined, the large mixed data set consists of 11588 samples of which 10493 were primary cancer samples, 715 were normal tissue samples, and 375 were metastatic cancer samples.

The mixed data set was annotated to include labels for the 4 different classification tasks within the model architecture: organ system of origin, disease state, cancer type, and cancer

subtype. A summary of the categories for the mixed data set can be found in Tables 2.2, 2.3, 2.4, 2.5, and 2.6. There are 3 labels for the disease state category corresponding to primary cancer, metastatic cancer, and normal tissue samples. The other classification tasks consist of 11 organ systems of origin, 68 cancer types, and 91 cancer subtypes. The number of classes presented here reflect those remaining after the preprocessing/filtering steps outlined in Section 2.1.2. Within both the cancer type and subtype labels, there are 20 metastatic cancers and 16 normal classes. Within the organ system of origin task there are 8 classes that have normal samples included.

Total Number of Cancer Subtypes	**91**
Number of Primary Subtypes	55
Number of Metastatic Subtypes	20
Number of Normal Subtypes	16
Total Number of Cancer Types	**68**
Number of Primary Subtypes	32
Number of Metastatic Types	20
Number of Normal Types	16
Total Number of Organ Systems of Origin	**11**
Number of Organ Systems with Normal Samples	8
Total Number of Tissue Types	**3**

Table 2.2: Number and composition of classes for each classification task

Organ System of Origin		
Full Name	*Number of Cancer Samples*	*Number of Normal Samples*
Breast	1268	112
Central Nervous System	1024	0
Endocrine	1005	59
Gastrointestinal	1756	146
Gynecologic	883	24

Organ System of Origin		
Full Name	*Number of Cancer Samples*	*Number of Normal Samples*
Head and Neck	650	44
Hematologic	615	0
Skin	484	0
Soft Tissue	294	0
Thoracic	1436	110
Urological	2071	200
Total Number of Cancer Samples	**10496**	
Total Number of Normal Samples		**695**
Total Number of Samples		**11486**

Table 2.3: Organ system of origin classes and frequencies within the full set of preprocessed data (including both train and test data)

Tissue Type	
Full Name	*Number of Samples*
Primary Tumour	10496
Metastatic Tumour	295
Normal Tissue	695
Total Number of Samples	**11486**

Table 2.4: Tissue type classes and frequencies within the full set of preprocessed data (including both train and test data)

Cancer Types		
Abbreviation	*Full Name*	*Number of Samples*
ACC_T_Metastatic	Metastatic Adrenocortical Carcinoma	8
ACC_T_Tumor	Adrenocortical Carcinoma	79

Cancer Types		
Abbreviation	*Full Name*	*Number of Samples*
ALL_T_Metastatic	Acute Lymphocytic Leukemia	13
BLCA_N_Normal	Bladder Tissue	19
BLCA_T_Metastatic	Metastatic Bladder Urothelial Carcinoma	14
BLCA_T_Tumor	Bladder Urothelial Carcinoma	408
BRCA_T_Metastatic	Metastatic Breast Invasive Carcinoma	56
BRCA_N_Normal	Breast Tissue	112
BRCA_T_Tumor	Breast Invasive Carcinoma	1100
CESC_T_Tumor	Endocervical Adenocarcinoma	300
CHOL_T_Metastatic	Extrahepatic Cholangiocarcinoma	19
CHOL_N_Normal	Bile Duct Tissue	9
CHOL_T_Tumor	Cholangiocarcinoma	36
CLL_T_Tumor	Chronic Lymphocytic Leukemia	29
CML_T_Tumor	Chronic Myelogenous Leukemia	102
COADREAD_N_Normal	Colorectal Tissue	51
COADREAD_T_Metastatic	Metastatic Colorectal Adenocarcinoma	10
COADREAD_T_Tumor	Colorectal Adenocarcinoma	386
DLBC_T_Tumor	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	170
ESCA_T_Metastatic	Metastatic Esophageal Adenocarcinoma	9
ESCA_T_Tumor	Esophageal Carcinoma	169
FL_T_Tumor	Follicular Lymphoma	50
GBM_T_Tumor	Glioblastoma	219
HNSC_N_Normal	Head and Neck Tissue	44
HNSC_T_Metastatic	Metastatic Head and Neck Squamous Cell Carcinoma	9

Cancer Types		
Abbreviation	*Full Name*	*Number of Samples*
HNSC_T_Tumor	Head and Neck Squamous Cell Carcinoma	517
KICH_N_Normal	Kidney Tissue	25
KICH_T_Tumor	Kidney Chromophobe	66
KIRC_N_Normal	Kidney Tissue	72
KIRC_T_Tumor	Kidney Renal Clear Cell Carcinoma	532
KIRP_N_Normal	Kidney Tissue	32
KIRP_T_Tumor	Kidney Renal Papillary Cell Carcinoma	291
LAML_T_Metastatic	Metastatic Acute Myeloid Leukemia	8
LAML_T_Tumor	Acute Myeloid Leukemia	123
LGG_T_Tumor	Brain Lower Grade Glioma	530
LIHC_N_Normal	Liver Tissue	50
LIHC_T_Metastatic	Metastatic Liver Hepatocellular Carcinoma	7
LIHC_T_Tumor	Liver Hepatocellular Carcinoma	373
LUAD_N_Normal	Lung Tissue	59
LUAD_T_Metastatic	Metastatic Lung Adenocarcinoma	9
LUAD_T_Tumor	Lung Adenocarcinoma	518
LUSC_N_Normal	Lung Tissue	51
LUSC_T_Tumor	Lung Squamous Cell Carcinoma	501
MB-Adult_T_Tumor	Medulloblastoma	275
MESO_T_Tumor	Mesothelioma	298
OV_T_Metastatic	Metastatic Ovarian Serous Cystadenocarcinoma	13
OV_T_Tumor	Ovarian Serous Cystadenocarcinoma	308
PAAD_T_Metastatic	Metastatic Pancreatic Adenocarcinoma	7

Cancer Types		
Abbreviation	*Full Name*	*Number of Samples*
PAAD_T_Tumor	Pancreatic Adenocarcinoma	179
PCPG_T_Tumor	Pheochromocytoma and Paraganglioma	184
PRAD_N_Normal	Prostate Tissue	52
PRAD_T_Metastatic	Metastatic Prostate Adenocarcinoma	62
PRAD_T_Tumor	Prostate Adenocarcinoma	498
NET_T_Metastatic	Metastatic Neuroendocrine Tumour	6
SARC_T_Metastatic	Metastatic Sarcoma	33
SARC_T_Tumor	Sarcoma	261
SKCM_T_Metastatic	Metastatic Skin Cutaneous Melanoma	12
SKCM_T_Tumor	Skin Cutaneous Melanoma	472
STAD_N_Normal	Stomach Tissue	36
STAD_T_Tumor	Stomach Adenocarcinoma	415
TGCT_T_Tumor	Testicular Germ Cell Tumors	156
THCA_N_Normal	Thyroid Tissue	59
THCA_T_Tumor	Thyroid Carcinoma	513
THYM_T_Tumor	Thymoma	120
UCEC_N_Normal	Uterine Tissue	24
UCEC_T_Tumor	Uterine Corpus Endometrial Carcinoma	181
UCS_T_Tumor	Uterine Carcinosarcoma	57
UVM_T_Tumor	Uveal Melanoma	80
Total Number of Primary Samples		**10496**
Total Number of Metastatic Samples		**295**
Total Number of Normal Samples		**695**
Total Number of Samples		**11486**

Cancer Types		
Abbreviation	*Full Name*	*Number of Samples*

Table 2.5: Cancer type class abbreviations and frequency within the full set of preprocessed data (including both train and test data)

Cancer Subtypes		
Abbreviation	*Full Name*	*Number of Samples*
ACC_T_Metastatic	Metastatic Adrenocortical Carcinoma	8
ACC_T_Tumor	Adrenocortical Carcinoma	79
ALL_T_Metastatic	Acute Lymphocytic Leukemia	13
BLCA_N_Normal	Bladder Tissue	19
BLCA_T_Metastatic	Metastatic Bladder Urothelial Carcinoma	14
BLCA_T_Tumor	Bladder Urothelial Carcinoma	408
BRCA_Basal_T_Tumor	Basal Breast Invasive Carcinoma	176
BRCA_HER2like_Tumor	HER2-like Breast Invasive Carcinoma	80
BRCA_IDC_T_Metastatic	Metastatic Invasive Ductal Breast Carcinoma	46
BRCA_ILC_T_Metastatic	Metastatic Invasive Lobular Breast Carcinoma	10
BRCA_LuminalA_T_Tumor	Luminal A Breast Invasive Carcinoma	538
BRCA_LuminalB_T_Tumor	Luminal B Breast Invasive Carcinoma	207
BRCA_N_Normal	Breast Tissue	112
BRCA_T_Tumor	Breast Invasive Carcinoma	99
CESC_CAD_T_Tumor	Endocervical Adenocarcinoma	47
CESC_SCC_T_Tumor	Cervical Squamous Cell Carcinoma and Endocervical Adenocarcinoma	253
CHOL_EHCH_T_Metastatic	Metastatic Extrahepatic Cholangiocarcinoma	10

Cancer Subtypes		
Abbreviation	*Full Name*	*Number of Samples*
CHOL_IHCH_T_Metastatic	Metastatic Intrahepatic Cholangiocarcinoma	9
CHOL_N_Normal	Bile Duct Tissue	9
CHOL_T_Tumor	Cholangiocarcinoma	36
CLL_T_Tumor	Chronic Lymphocytic Leukemia	29
CML_T_Tumor	Chronic Myelogenous Leukemia	102
COADREAD_N_Normal	Colorectal Tissue	51
COADREAD_T_Metastatic	Metastatic Colorectal Adenocarcinoma	10
COADREAD_T_Tumor	Colorectal Adenocarcinoma	386
DLBC_BM_T_Tumor	Bone Marrow Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	11
DLBC_T_Tumor	Lymphoid Neoplasm Diffuse Large B-cell Lymphoma	159
ESCA_EAC_T_Metastatic	Metastatic Esophageal Adenocarcinoma	9
ESCA_EAC_T_Tumor	Esophageal Adenocarcinoma	63
ESCA_SCC_T_Tumor	Squamous Cell Esophageal Carcinoma	93
ESCA_T_Tumor	Esophageal Carcinoma	13
FL_T_Tumor	Follicular Lymphoma	50
GBM_T_Tumor	Glioblastoma	219
HNSC_N_Normal	Head and Neck Tissue	44
HNSC_T_Metastatic	Metastatic Head and Neck Squamous Cell Carcinoma	9
HNSC_T_Tumor	Head and Neck Squamous Cell Carcinoma	517
KICH_N_Normal	Kidney Tissue	25
KICH_T_Tumor	Kidney Chromophobe	66

Cancer Subtypes		
Abbreviation	*Full Name*	*Number of Samples*
KIRC_N_Normal	Kidney Tissue	72
KIRC_T_Tumor	Kidney Renal Clear Cell Carcinoma	532
KIRP_N_Normal	Kidney Tissue	32
KIRP_T_Tumor	Kidney Renal Papillary Cell Carcinoma	291
LAML_T_Metastatic	Metastatic Acute Myeloid Leukemia	8
LAML_T_Tumor	Acute Myeloid Leukemia	123
LGG_T_Tumor	Brain Lower Grade Glioma	530
LIHC_N_Normal	Liver Tissue	50
LIHC_T_Metastatic	Metastatic Liver Hepatocellular Carcinoma	7
LIHC_T_Tumor	Liver Hepatocellular Carcinoma	373
LUAD_N_Normal	Lung Tissue	59
LUAD_T_Metastatic	Metastatic Lung Adenocarcinoma	9
LUAD_T_Tumor	Lung Adenocarcinoma	518
LUSC_N_Normal	Lung Tissue	51
LUSC_T_Tumor	Lung Squamous Cell Carcinoma	501
MB_Group3_T_Tumor	Group 3 Medulloblastoma	39
MB_Group4_T_Tumor	Group 4 Medulloblastoma	69
MB_SHH_T_Tumor	Sonic Hedgehog Medulloblastoma	136
MB_WNT_T_Tumor	Wingless Medulloblastoma	31
MESO_T_Tumor	Mesothelioma	298
OV_T_Metastatic	Metastatic Ovarian Serous Cystadenocarcinoma	13
OV_T_Tumor	Ovarian Serous Cystadenocarcinoma	308
PAAD_T_Metastatic	Metastatic Pancreatic Adenocarcinoma	7
PAAD_T_Tumor	Pancreatic Adenocarcinoma	179

Cancer Subtypes		
Abbreviation	*Full Name*	*Number of Samples*
PCPG_T_Tumor	Pheochromocytoma and Paraganglioma	184
PRAD_N_Normal	Prostate Tissue	52
PRAD_T_Metastatic	Metastatic Prostate Adenocarcinoma	62
PRAD_T_Tumor	Prostate Adenocarcinoma	498
PrNET_T_Metastatic	Metastatic Pancreatic Neuroendocrine Tumour	6
SARC_DDL_T_Tumor	Dedifferentiated Sarcoma	58
SARC_LMS_T_Metastatic	Leiomyosarcoma	9
SARC_LMS_T_Tumor	Dedifferentiated Liposarcoma	106
SARC_MPNST_T_Tumor	Malignant Peripheral Nerve Sheath Tumour	10
SKCM_T_Metastatic	Metastatic Skin Cutaneous Melanoma	12
SKCM_T_Tumor	Skin Cutaneous Melanoma	472
STAD_CIN_T_Tumor	Chromosomal Instability Stomach Adenocarcinoma	211
STAD_EBV_T_Tumor	EBV-positive Stomach Adenocarcinoma	31
STAD_GS_T_Tumor	Genomically Stable Stomach Adenocarcinoma	70
STAD_MSI_T_Tumor	Microsatellite Instability Stomach Adenocarcinoma	76
STAD_N_Normal	Stomach Tissue	36
STAD_T_Tumor	Stomach Adenocarcinoma	27
TGCT_T_Tumor	Testicular Germ Cell Tumors	156
THCA_N_Normal	Thyroid Tissue	59
THCA_T_Tumor	Thyroid Carcinoma	513

Cancer Subtypes		
Abbreviation	*Full Name*	*Number of Samples*
THYM_T_Tumor	Thymoma	120
UCEC_N_Normal	Uterine Tissue	24
UCEC_T_Tumor	Uterine Corpus Endometrial Carcinoma	181
UCS_T_Tumor	Uterine Carcinosarcoma	57
UVM_T_Tumor	Uveal Melanoma	80
Total Number of Primary Samples		**10496**
Total Number of Metastatic Samples		**295**
Total Number of Normal Samples		**695**
Total Number of Samples		**11486**

Table 2.6: Cancer subtype class abbreviations and frequency within the full set of preprocessed data (including both train and test data)

Metastatic-Only Data Set

The second data set was derived from the Personalised OncoGenomics (POG) project at BC Cancer and contains only metastatic cancer samples. Throughout this thesis, this data set is referred to as the external test set, the POG data set, or the metastatic-only test set. Extensive details of the POG project can be found in the paper by Pleasance et al. [**60**]. A summary of its composition as utilized in this thesis can be found in Tables 2.7, 2.8, and 2.9. There are 461 metastatic cancer samples that span 15 cancer subtypes, 13 cancer types, and 10 organ systems of origin. The 461 samples were selected from a larger set of POG data and were chosen on the basis that each of their labels for all four classification tasks were also present in the training data.

Organ System of Origin	
Full Name	*Number of Samples*
Breast	134
Endocrine	6
Gastrointestinal	163
Gynecologic	34
Head and Neck	5
Hematologic	2
Skin	14
Soft Tissue	56
Thoracic	44
Urological	3
Total Number of Organ Systems	**11**
Total Number of Samples	**461**

Table 2.7: Organ system of origin classes and frequencies within the the POG dataset

Cancer Types		
Abbreviation	*Full Name*	*Number of Samples*
ACC_T_Metastatic	Metastatic Adrenocortical Carcinoma	6
BRCA_T_Metastatic	Metastatic Breast Invasive Carcinoma	134
CHOL_T_Metastatic	Metastatic Cholangiocarcinoma	3
COADREAD_T_Metastatic	Metastatic Colorectal Adenocarcinoma	85
HNSC_T_Metastatic	Metastatic Head and Neck Squamous Cell Carcinoma	5
LAML_T_Metastatic	Metastatic Acute Myeloid Leukemia	2
LIHC_T_Metastatic	Metastatic Liver Hepatocellular Carcinoma	3
LUAD_T_Metastatic	Metastatic Lung Adenocarcinoma	44

Cancer Types		
Abbreviation	*Full Name*	*Number of Samples*
OV_T_Metastatic	Metastatic Ovarian Serous Cystadenocarcinoma	34
PAAD_T_Metastatic	Metastatic Pancreatic Adenocarcinoma	72
PRAD_T_Metastatic	Metastatic Prostate Adenocarcinoma	3
SARC_T_Metastatic	Metastatic Sarcoma	56
SKCM_T_Metastatic	Metastatic Skin Cutaneous Melanoma	14
Total Number of Cancer Types		**13**
Total Number of Samples		**461**

Table 2.8: Cancer type class abbreviations and frequency within the POG dataset

Cancer Subtypes		
Abbreviation	*Full Name*	*Number of Samples*
ACC_T_Metastatic	Metastatic Adrenocortical Carcinoma	6
BRCA_IDC_T_Metastatic	Metastatic Invasive Ductal Breast Carcinoma	125
BRCA_ILC_T_Metastatic	Metastatic Invasive Lobular Breast Carcinoma	9
CHOL_IHCH_T_Metastatic	Metastatic Intrahepatic Cholangiocarcinoma	3
COADREAD_T_Metastatic	Metastatic Colorectal Adenocarcinoma	85
HNSC_T_Metastatic	Metastatic Head and Neck Squamous Cell Carcinoma	5
LAML_T_Metastatic	Metastatic Acute Myeloid Leukemia	2
LIHC_T_Metastatic	Metastatic Liver Hepatocellular Carcinoma	3
LUAD_T_Metastatic	Metastatic Lung Adenocarcinoma	44

Cancer Subtypes		
Abbreviation	*Full Name*	*Number of Samples*
OV_T_Metastatic	Metastatic Ovarian Serous Cystadenocarcinoma	34
PAAD_T_Metastatic	Metastatic Pancreatic Adenocarcinoma	72
PRAD_T_Metastatic	Metastatic Prostate Adenocarcinoma	3
SARC_LMS_T_Metastatic	Metastatic Leiomyosarcoma	11
SARC_T_Metastatic	Metastatic Sarcoma	45
SKCM_T_Metastatic	Metastatic Skin Cutaneous Melanoma	14
Total Number of Cancer Subtypes		**15**
Total Number of Samples		**461**

Table 2.9: Cancer subtype class abbreviations and frequency within the POG dataset

The POG data class labels were annotated using the same class labels that were found in the training data and correspond to the TCGA naming convention. The most appropriate TCGA label was determined as part of the analysis conducted for the POG project and considered genomic, pathological, and clinical factors [**61**, **62**].

2.1.1 Training and Test Sets

The mixed held-out data set described above was divided into training and test sets. The training set used to train the model(s) was generated by utilizing 85% of the whole mixed data set and contained 9763 samples. The remaining 15%, 1723 samples, constitutes the held-out test data set and contains primary, metastatic, and normal samples in proportions equal to those found in the training data set (ie. it is stratified). This held-out data was excluded from all aspects of training including cross-validation.

The POG data as described above was utilized in its entirety for testing only. The resulting test set is 461 metastatic cancer samples. The value of this data set as a test set is that all of its samples were processed at a facility that is different from any of the metastatic samples in the training and held-out test set. This should add some objectivity to the testing results.

2.1.2 Data Preprocessing: Validation

The RPKM values of the data were ranked and normalized to lie between 0 and 1 using the *rank* function from the *pandas* Python package [**63**]. Samples were then filtered out based on whether or not they were part of a cancer subtype class that contained at least 6 samples. Since the intention was to utilize five-fold cross-validation for model validation and optimization, it was important to keep this minimum number of samples to ensure class ratios remained the same across all folds.

Following the filtering of samples, 15% of the data was separated into a held-out test set using the *train_test_split* function found in the *scikit-learn* Python package [**64**]. The option to stratify the classes was enabled to ensure proper class representation. The remaining 85% of the data not used for the held-out test set was then divided into five folds for use in cross-validation. The *StratifiedKFold* function from the *scikit-learn* package was used to generate the folds and maintain class ratios. The result of the data splitting is that at least one sample of each subtype was present in the test set and five samples were equally split among the five folds generated for cross-validation.

2.1.3 Data Preprocessing: Testing

The preprocessing steps for testing differ slightly from those of validation outlined in Section 2.1.2. Since the model has been validated using cross-validation, multiple training folds is no longer necessary for training the final model. The advantage of this is that more data can be used for training the model as a validation set is not needed. Therefore, the

preprocessing for testing excludes splitting the data for cross-validation but still includes normalization and ranking via the *rank* function within the *pandas* Python package. The result is a held-out test set containing 15% of the mixed data and a single training data set containing the remaining 85%. The metastatic-only test data (derived from POG) underwent the same ranking and normalization described above.

2.2 The Model

The model used in this thesis is a fully-connected feed-forward artificial neural network. The model is a multi-task model in that it has four classification output layers used to make classifications within four distinct tasks. A visualization of the model can be seen in Figure 2.1. These four classification tasks are:

1. Organ System of Origin
2. Disease State
3. Cancer Type
4. Cancer Subtype

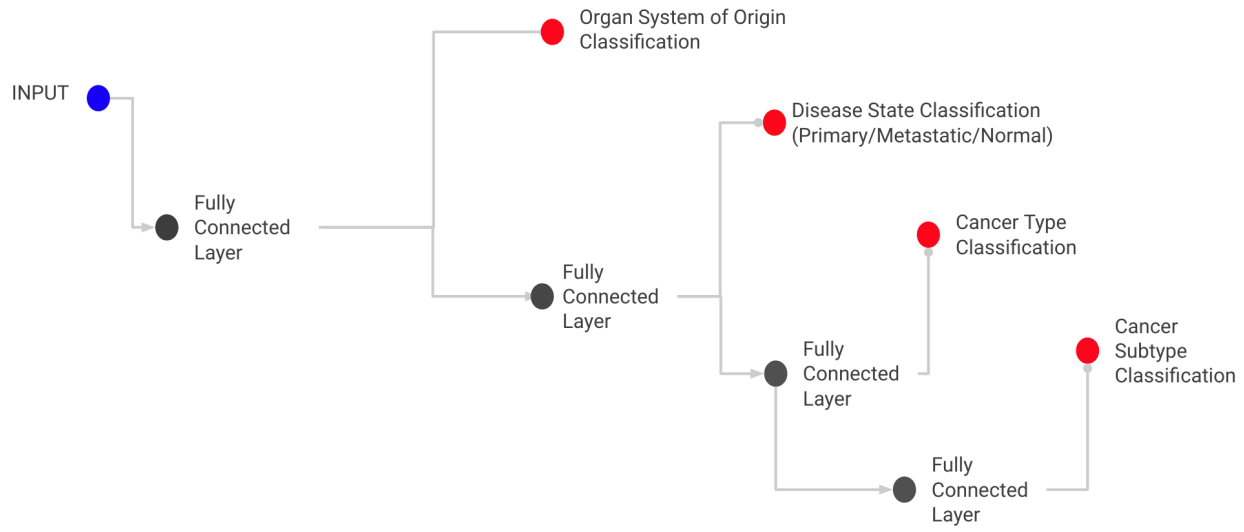

Figure 2.1: High level diagram of the multi-task neural network

Each classification task in the neural network model has a hidden layer directly connected with it. Each hidden layer connects to a classification task output layer as well as the next hidden layer (with the exception of the final hidden layer). Each hidden layer is a fully connected (dense) layer. The effect of this network architecture is that as information moves from the first hidden layer (associated with the organ system of origin classification) to the final hidden layer (associated with the cancer subtype classification), the model has more hidden layers to utilize in making the classification. For example, with the organ system of origin classification there is only a single hidden layer available to encode information, but at the cancer subtype classification there are four. By having more hidden layers for learning tasks that are more complex (cancer subtype being more complex than organ system), we are providing the model with a greater learning capacity for these more complex tasks.

The rationale behind using a model with this multi-task architecture is four fold. These are described below.

The first rationale is that the multiple task setup forces the model to learn increasing granular features of the data. The first dense layer must encode all of the information

necessary to accurately classify an organ system of origin. The effect of this is that in subsequent layers (down-stream from the organ system layer), the model is encouraged to learn features that will help to distinguish the disease state, and can, at least in part, ignore features needed to distinguish the organ system of origin.

The increasing granulation of feature learning described above contributes to the second rationale behind this architecture: mitigating tissue bias. A single learning task that requires a model to only classify cancer types and seeks to do so with cancers from different organ systems will, at least in part, learn features that define the organ system. This is tissue bias. We can imagine trying to distinguish stomach cancer from brain cancer. During training, the model can increase its baseline classification accuracy if it can learn what makes a brain different from a stomach. This does not necessarily require learning anything about cancer specifically. The background gene expression levels of the relevant organ systems can be leveraged in distinguishing brain cancer from stomach cancer and may be enough information to accurately classifying some samples. Thus, the model is encouraged to identify the expression patterns of the organ system of origin. By forcing the model to learn to distinguish organ systems with the first layer of the multi-task model, we are providing a mechanism of encouragement for it to learn patterns of expression specific to cancers in subsequent learning tasks.

The third rationale behind this multi-task architecture is that we are imbuing a biological hierarchy into the decision making process. The order of classification tasks is such that it follows a biological hierarchy. The model first questions what organ system is involved, then if this is normal or cancerous tissue, then what type of cancer it is, followed by what subtype. This is a biologically relevant series of decisions and may help to improve the classification ability of the model. In fact, convolutional neural networks used for image recognition are thought to show improved performance as a result of the hierarchical nature of their structure and learning [**65**]. It is reasonable to attempt to utilize this approach in the domain of this thesis work.

The fourth and final rationale for this multi-task model is simply the volume of output. The more tasks we have, the greater the wealth of data being output. This is an advantage when conducting post-classification analysis of the trained model as it provides access to more decision levels of the model and may provide a means with which to ask more interesting biological questions.

2.2.1 Model Settings and Hyperparameters

The optimization of the model's hyperparameters was done using five-fold cross-validation and a combination of manual search and limited grid search. The performance on the mean of all five validation sets was examined to determine the hyperparameter values of the model. The hyperparameters experimented with included the number of nodes in the hidden layers of the model, the learning rate, optimizer, dropout rate, batch sizes and various learning rate decay schedules. Ultimately, the hyperparameter settings seen in Table 2.10 were the best values found for this particular model architecture and problem space.

Hyperparameter	**Value**
Number of Nodes in Hidden Layers	2000
Optimizer	Mini-Batch Gradient Descent
Batch size	32
Learning Rate (reduce on plateau)	0.001
Learning Rate Reduction Factor	0.95
Learning Rate Reduction Patience	20
Early Stopping Patience	40
Dense Layer Activation Function	Tanh
Dropout (every dense layer)	0.2
Class Weighting	True

Table 2.10: Hyperparameter settings

Initilization

The weights were initialized using the glorot uniform (also known as the Xavier uniform) initializer as implemented in the *Keras* Python package [**51**]. The biases were initialized to zeros.

Mini-Batch Gradient Descent

Mini-batch gradient descent was used as the optimizer for the model. The implementation used was the one found in the *Keras* Python package [**51**]. This is simply the *SGD* optimizer with the batch size set to 32.

Learning Rate Reduction

For this work, the validation loss on the cancer subtype classification task was used as the observed metric for learning rate reduction patience. The initial learning rate was set to 0.001, the reduction factor to 0.95, and the reduction patience to 20 (Table 2.10). The learning rate reduction was implemented using the *ReduceLROnPlateau* callback from the *Keras* Python package [**51**].

Early Stopping and Patience

When determining if training should be stopped at any given epoch due to a lack of improvement in the validation loss, a patience of 40 was utilized. This means that the model would allow 40 epochs to complete without an improvement in the validation loss before halting the training. The *EarlyStopping* callback from the *Keras* Python package was utilized to achieve early stopping and patience for the models used in this thesis [**51**].

Activation Function

The hyperbolic tangent function was used as the activation function for each node within the hidden layers of all of the models presented in this thesis. The softmax activation function was used for each node in the output layers. Both activation functions were used as implemented in the *Keras* Python package [**51**].

Loss Function

The categorical cross-entropy loss was utilized for the models in this thesis. The implementation used was the standard one found in the *Keras* Python package [**51**].

Dropout

A dropout rate of 0.2 or 20% was used for the models in this thesis and was implemented using the *Dropout* layers from the *Keras* Python package [**51**].

Class Weighting

Class weighting was implemented for the models in this thesis using the *compute_class_weight* and *compute_sample_weight* functions from the *scikit-learn* Python package [**16**]. Due to limitations of the *Keras* package in a multi-task environment, it was not possible to directly apply the class weights during training. As a workaround, the *compute_class_weight* function was used to calculate the proper weight values and then they were applied on a per-sample level at training time using the *compute_sample_weight* function. The weights were chosen to reflect the relative sizes of the classes. The largest class was given a weight of 1 and all other classes were given a weight corresponding to the difference in their sizes compared to the largest class. For example, if a minority class had half the number of samples as the majority class, it was assigned a weight of 2.

2.2.2 Evaluating the Effect of Multiple Tasks on Classification Performance

The classification performance of the multi-task model was evaluated against models containing fewer learning tasks and one model using just a single task (cancer subtype only). The evaluation of the models was conducted as part of the cross-validation process and thus the results presented here are an average of the performance across five validation folds. The validation folds contained normal tissue, primary cancer, and metastatic cancer samples as described in Section 2.1. The validation folds were used to ensure that the test sets remained untouched during the validation stage. The following sections will present the validation results for each of the learning tasks: organ system of origin, disease state, cancer type, and cancer subtype.

Organ System of Origin

The F1-scores presented in Figure 2.2 range from 0.981639 for the "All Tasks" model to 0.984891 for a multi-task model containing organ system of origin, disease state, and cancer subtype and not containing a cancer type learning task ("No Cancer Type"). This represents a performance reduction of 0.003252 for the "All Tasks" model versus the best performing set of tasks. The variation in performance seen between models is largest between the "All Tasks" model and the other three models. We note that the smallest standard deviation is seen with the "No Disease State" model and the largest with the "No Cancer Type" model .

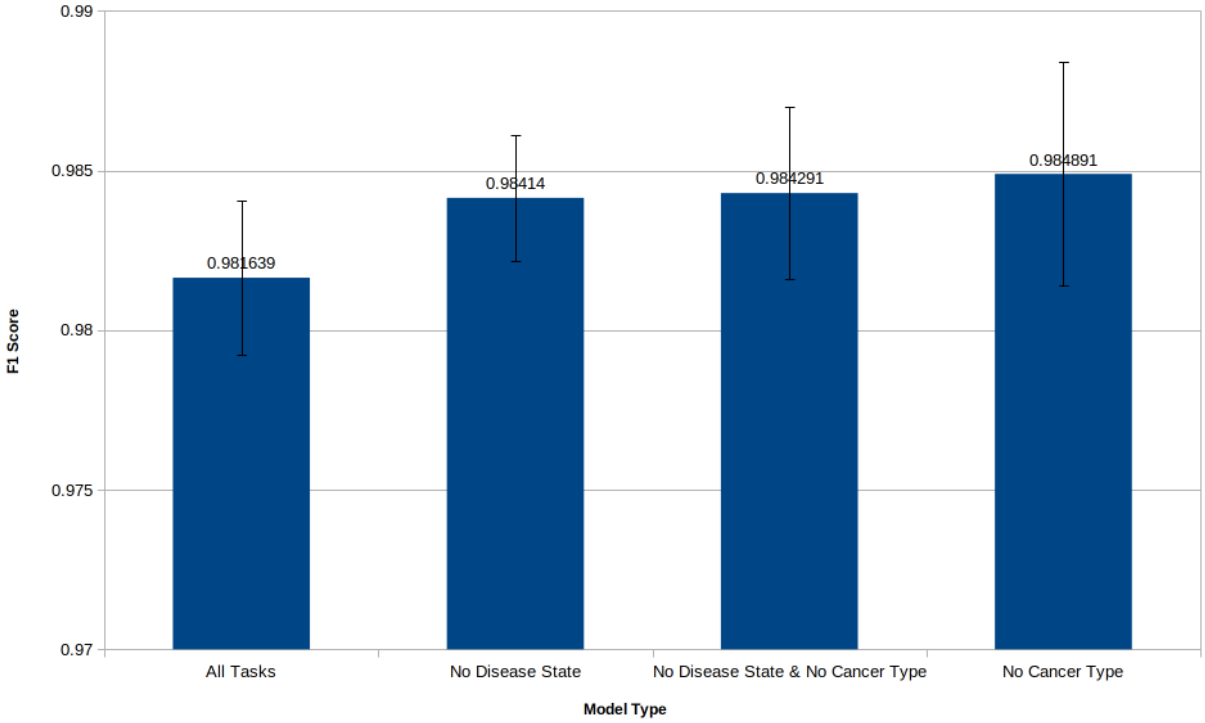

Figure 2.2: The macro F1-scores of various models using validation sets containing both primary and metastatic samples from different organ systems of origin.

Disease State

The F1-scores presented in Figure 2.3 range from 0.978703 for a model with the cancer type task removed and 0.981184 for a model missing the organ system of origin learning task ("No Organ System"). This represents a performance reduction of 0.002481. The performance of the "All Tasks" model sits in between the other two with an F1-score of 0.979349. The standard deviation is largest with the "No Cancer Type" model and smallest with the "No Organ System" model.

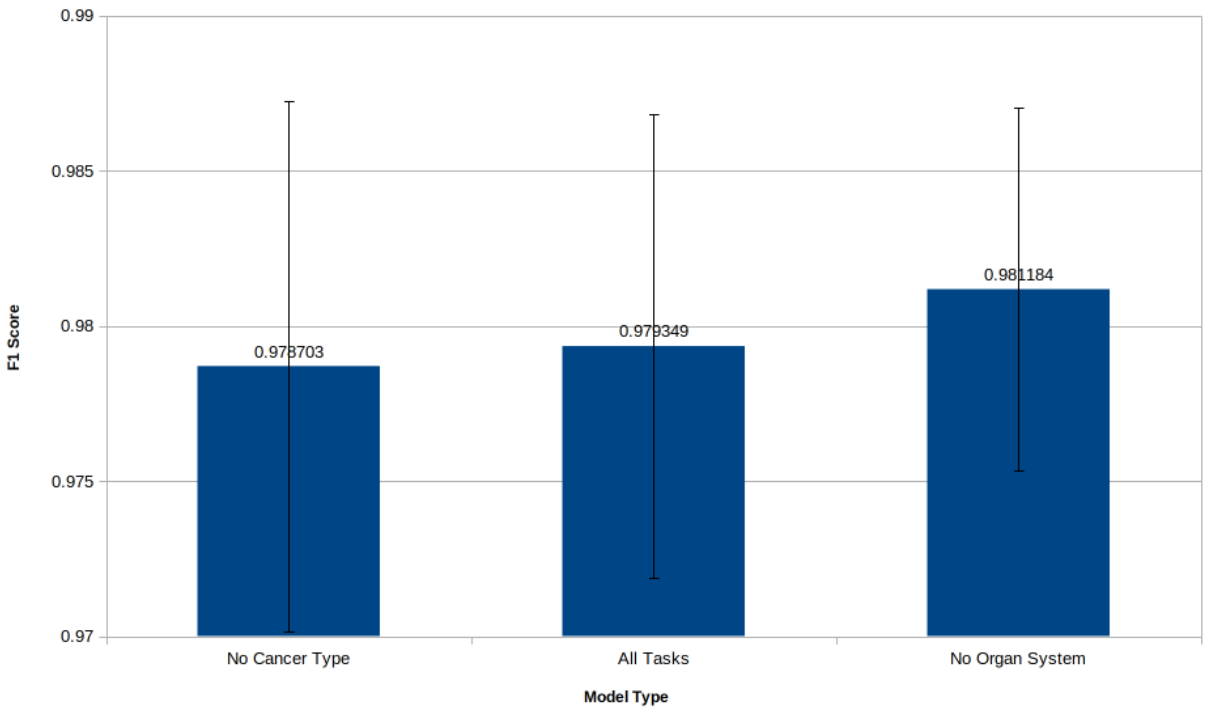

Figure 2.3: The macro F1-scores of various models on validation sets containing both primary and metastatic samples at the disease state classification level

Cancer Type

The F1-scores presented in Figure 2.4 range from 0.861412 for the "All Tasks" model to 0.862186 for a multi-task model containing organ system of origin, cancer type, and cancer subtype, and not containing a disease state learning task ("No Disease State"). This represents a performance improvement of 0.000774 over the "All Tasks" model, which had the poorest performance of the models tested. Note, however, that the "All Tasks" model had the smallest standard deviation. The variation in performance seen between models is much smaller for cancer type classification when compared with the cancer subtype classifications.

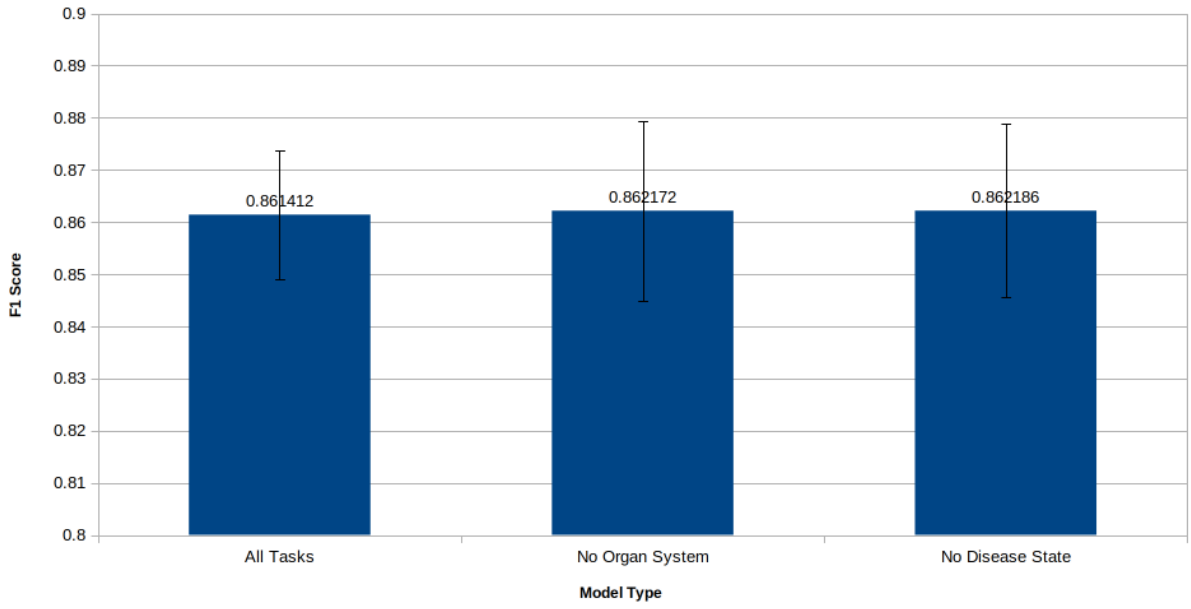

Figure 2.4: The macro F1-scores of various models on validation set data containing both primary and metastatic cancer type samples.

Cancer Subtype

The F1-scores presented in Figure 2.5 range from 0.80281 for the single task ("Subtype Only") model to 0.812746 for a multi-task model missing the disease state learning task ("No Disease State"). This represents a performance improvement of 0.009936 over the single task model. The performance of the "All Tasks" model is approximately in the middle of the other models with an F1-score of 0.806287. The performance decrease between the "All Tasks" model and the best performing model is 0.006459. Note that the "All Tasks" model had the largest standard deviation and the "No Cancer Type" model had the smallest.

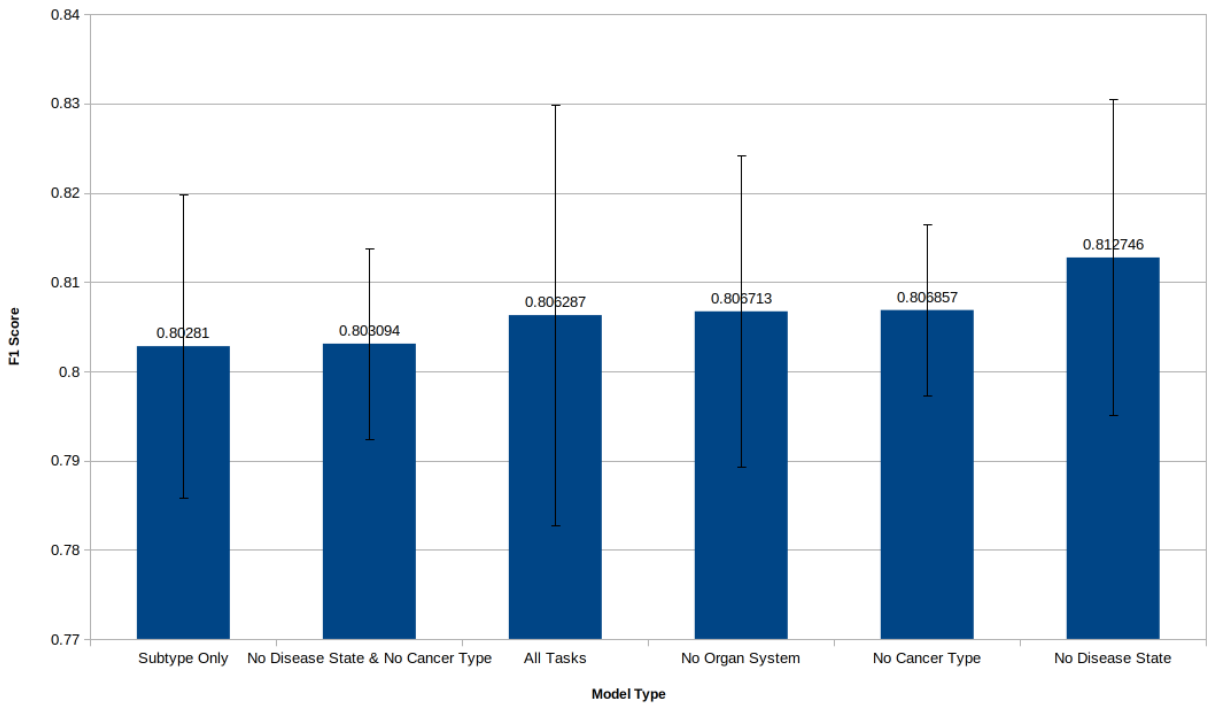

Figure 2.5: The macro F1-scores of various models on validation set data containing both primary and metastatic cancer subtype samples.

Discussion of Task-Dependent Performance

While the "All Tasks" model did not provide the best performance in any of the classification categories, relatively speaking, the difference in performance was small. In each classification category, the "All Tasks" model's mean performance was within an F1-score of 0.001 of the best performing set of tasks. The inclusion of all four tasks within the "All Tasks" model provides a greater opportunity to leverage more fine-grained information in down-stream analyses than it would if tasks were removed. Given that the classification performance is similar between all sets of tasks, it can be justified that the additional information gained from including all of the learning tasks is worth the slight loss of potential performance, and thus the "All Tasks" model can be used for further analysis.

Chapter 3

Classification of Cancers from Transcriptome Data

The following results were obtained from a single trained model. The first section presents the results from the mixed primary, metastatic and normal data that comprises the held-out test set as described in Chapter 2. The second section presents results using the external metastatic-only data set derived from POG data. Where F1-score is reported, it is the macro F1-score in which each class is weighted equally in the calculation of the score regardless of class size.

3.1 Results: Mixed Held-Out Test Set

3.1.1 Organ System of Origin

The organ system of origin classification scores can be seen in Table 3.2 and graphically in Figure 3.1. The classification performance was above 0.95 for all organ systems with the poorest performer being the soft tissue class. The soft tissue class was misclassified as thoracic and gastrointestinal at rate of approximately 2% and 3% respectively. The

misclassification of organ systems can be seen in Figure 3.2.

Organ System of Origin	**Precision**	**Recall**	**F1-score**	**Support**
Breast	1	1	1	190
Central Nervous System	1	1	1	154
Endocrine	0.993	0.993	0.993	151
Gastrointestinal	0.996	0.992	0.994	261
Gynecologic	0.97	0.985	0.978	133
Head and Neck	0.98	0.99	0.985	98
Hematologic	0.978	1	0.989	91
Skin	1	0.986	0.993	73
Soft Tissue	0.976	0.932	0.953	44
Thoracic	0.986	0.986	0.986	216
Urologic	0.987	0.984	0.986	312
accuracy	0.99	0.99	0.99	0.99
macro avg	0.988	0.986	0.987	1723
weighted avg	0.99	0.99	0.99	1723

Table 3.2: The precision, recall, F1-score, and support for each organ system of origin class with testing conducted using the mixed held-out test set.

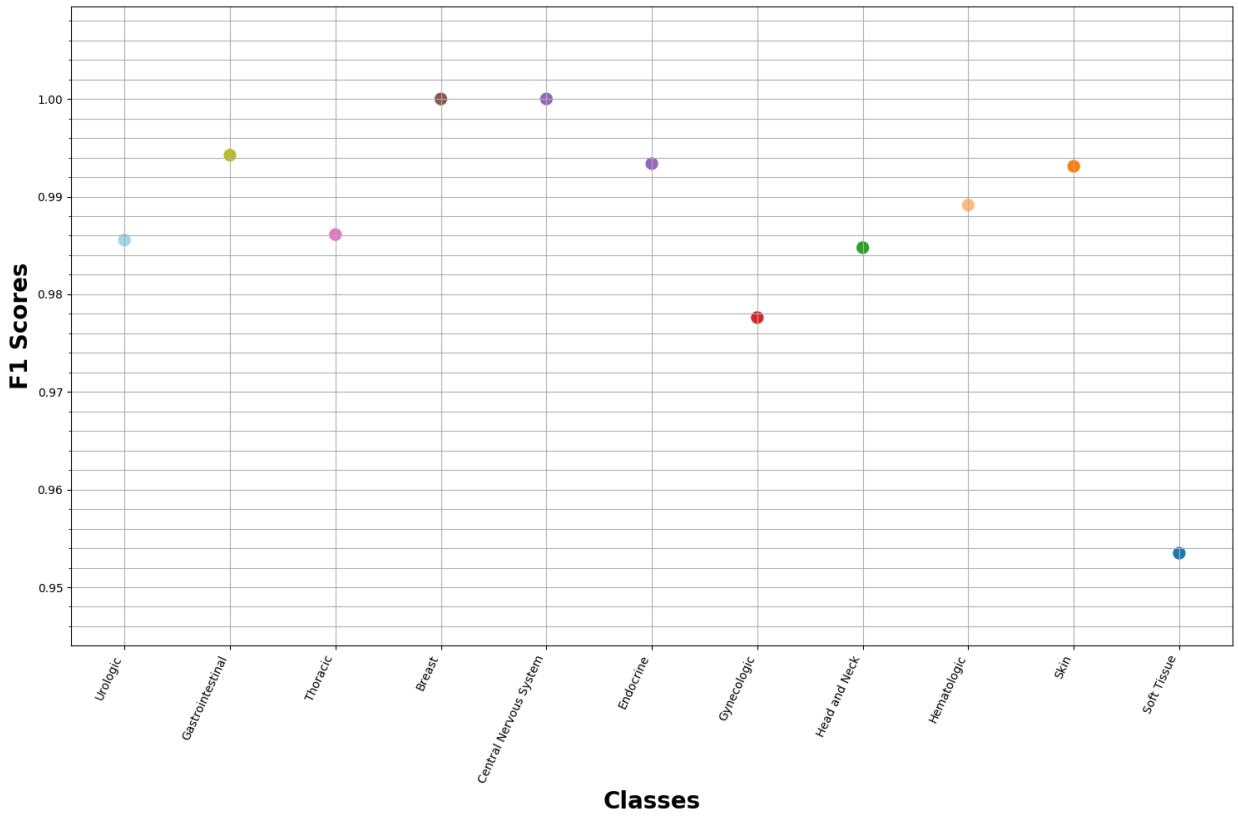

Figure 3.1: The macro F1-scores of each organ system of origin when testing on the held-out test set containing primary cancer, metastatic cancer, and normal tissue samples. Classes are ordered from left to right by the number of training samples available with colours representing bins of 20 samples.

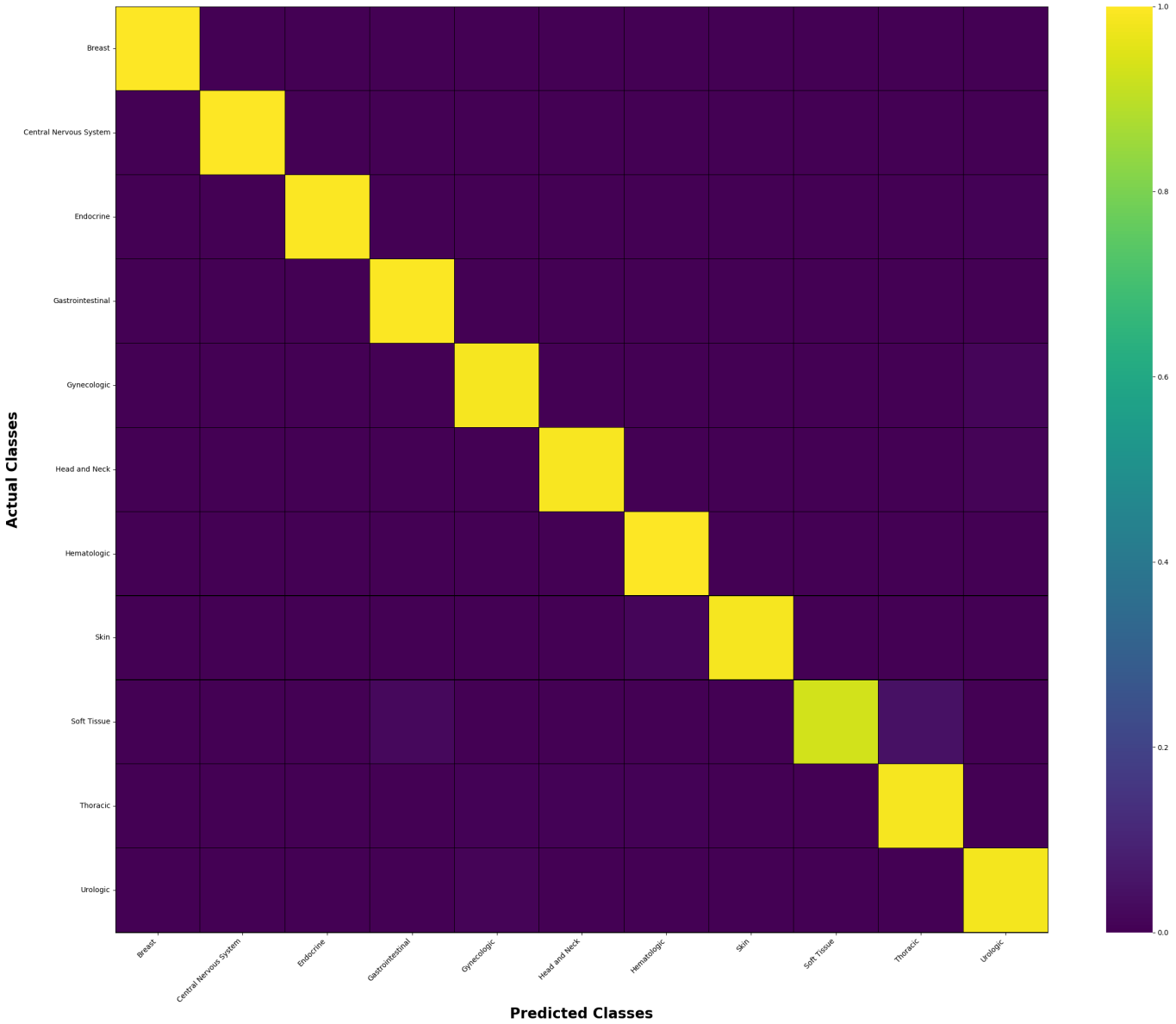

Figure 3.2: A confusion matrix depicting the organ system of origin classification performance on the held-out test set containing primary cancer, metastatic cancer, and normal tissue samples.

3.1.2 Disease State

The disease state classification scores can be seen in Table 3.4 and graphically in Figure 3.3. Each disease state had F1-scores above 0.95 with the poorest performer being the normal

class. Referring to Figure 3.4, we can see that the normal class was misclassified as the primary cancer class at rate of approximately 5%.

Disease State	**Precision**	**Recall**	**F1-score**	**Support**
Metastatic	1	1	1	41
Normal	0.98	0.934	0.957	106
Primary	0.996	0.999	0.997	1576
accuracy	0.995	0.995	0.995	0.995
macro avg	0.992	0.978	0.985	1723
weighted avg	0.995	0.995	0.995	1723

Table 3.4: The precision, recall, F1-score, and support for each disease state class with testing conducted using the mixed held-out test set.

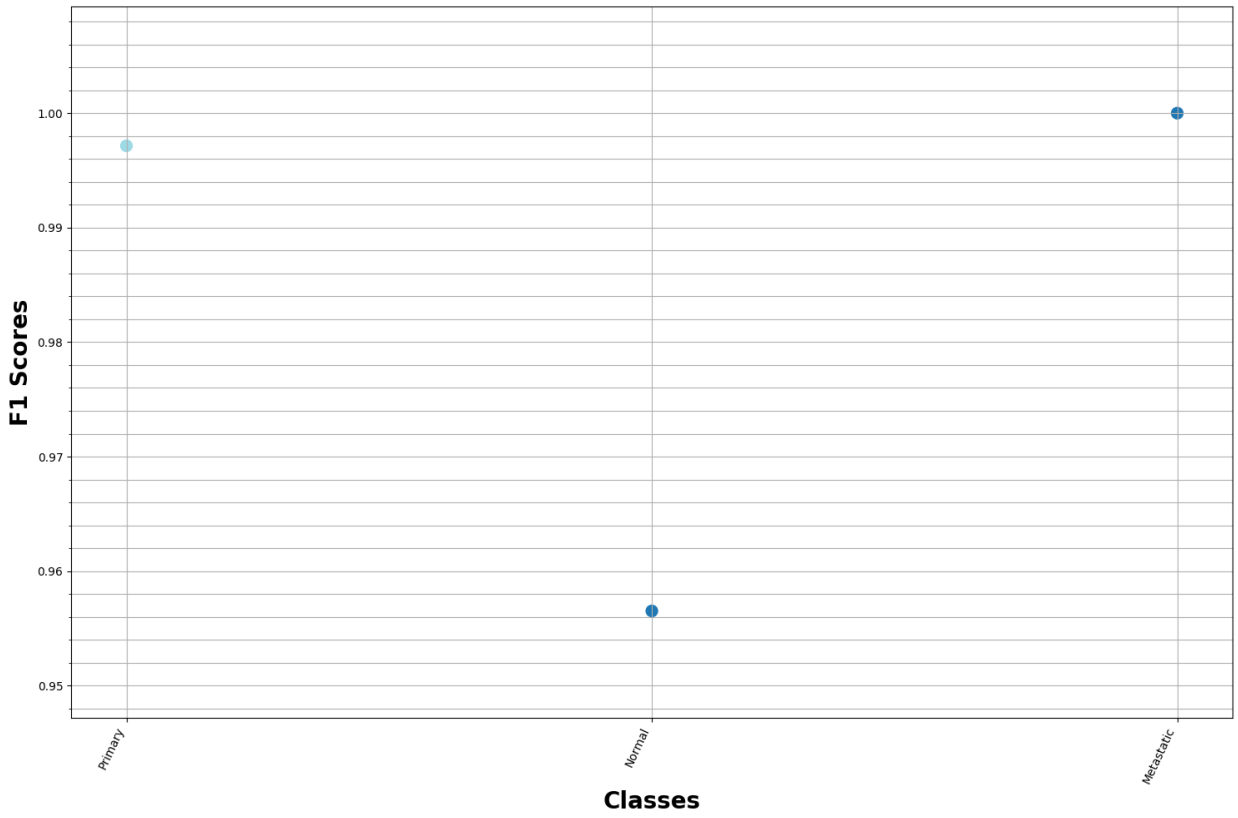

Figure 3.3: The macro F1-scores of each disease state when testing on the held-out test set containing primary cancer, metastatic cancer, and normal tissue samples. Classes are ordered from left to right by the number of training samples available with colours representing bins of 20 samples.

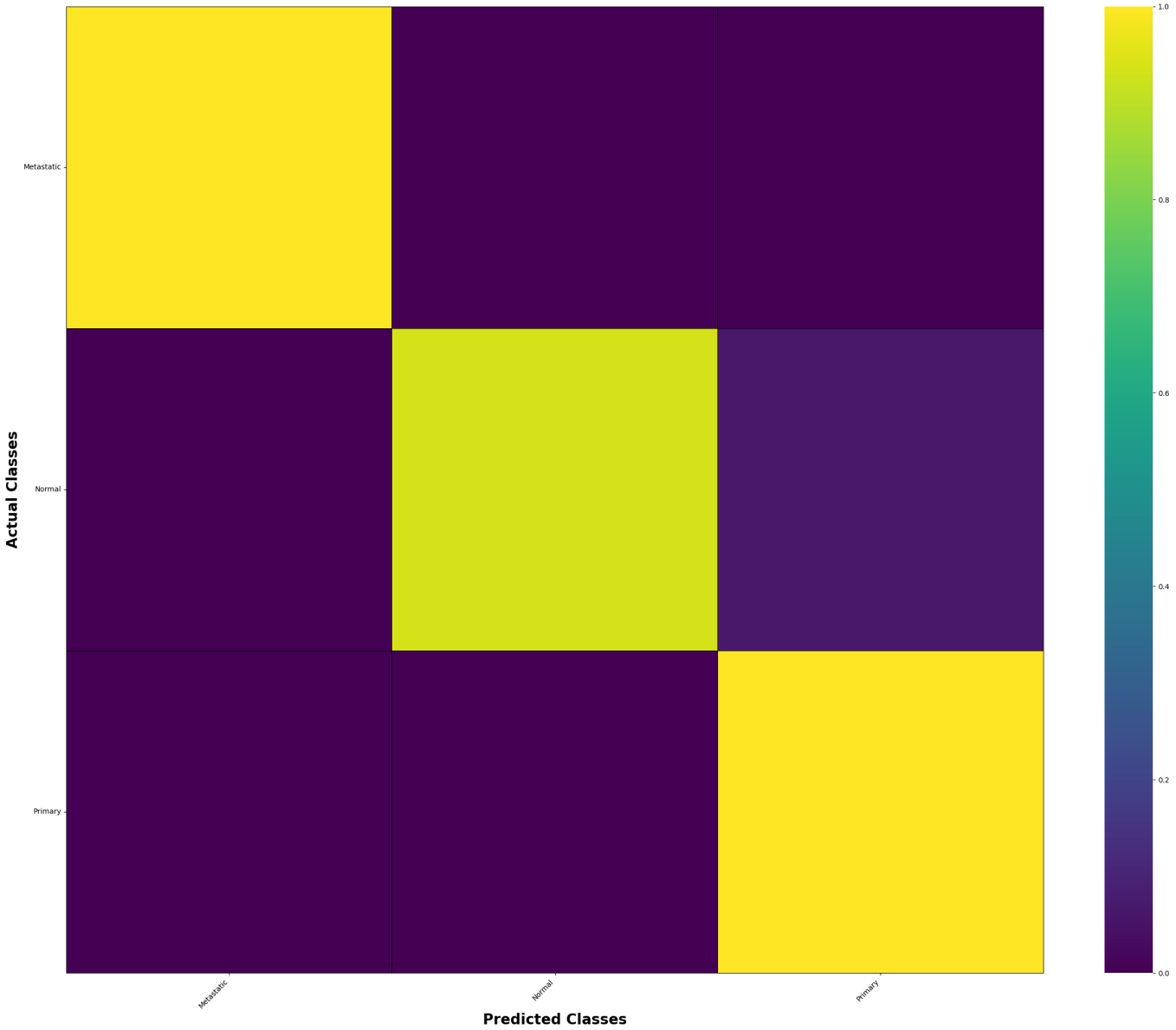

Figure 3.4: A confusion matrix depicting the disease state classification performance on the held-out test set containing primary cancer, metastatic cancer, and normal tissue samples.

3.1.3 Cancer Type

The classification performance of the cancer type task resulted in a total F1-score of 0.885 and an accuracy of 96.5% across all 68 types. The F1-scores for the classification of the primary, metastatic, and normal types individually were 0.964, 0.683, and 0.925 respectively

(see Figure 3.5).

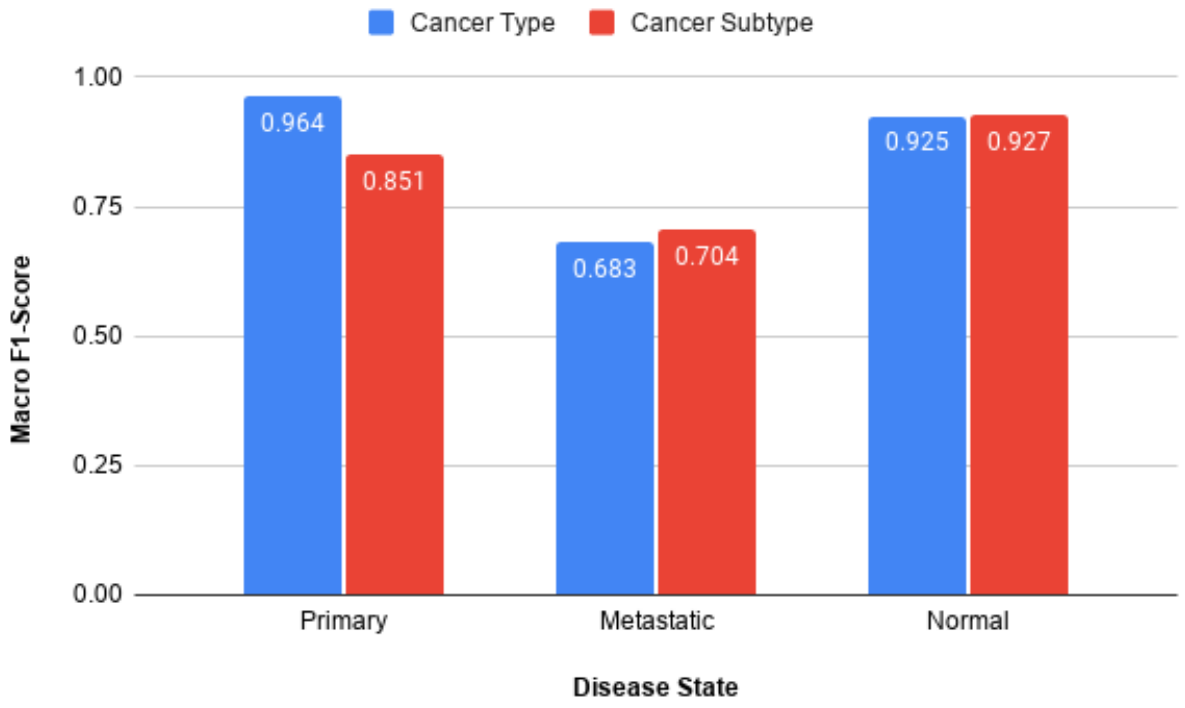

Figure 3.5: The macro F1-scores comparing the classification performance of cancer type and cancer subtype samples broken down by disease state as tested on the held-out mixed test set.

The F1-scores, precision, and recall of each cancer type can be seen in Table 3.6, and the class-wise F1-scores are presented graphically in Figure 3.6. The classification performance decreases along with the number of training samples. There are no outliers from the trend observed in Figure 3.6 like we saw with the subtype classes. However, there are four cancer types with an F1-score of 0.0. The cancer type classification accuracy and the predicted classes of misclassified types can be seen in the confusion matrix in Figure 3.9. The four cancer types with F1-scores of 0 and the source of their misclassifications are presented below:

ESCA_T_Metastatic

The ESCA_T_Metastatic type was completely misclassified as CHOL_T_Metastatic. Note that only a single test sample was available for this class.

HNSC_T_Metastatic

The HNSC_T_Metastatic type was completely misclassified as PAAD_T_Metastatic. Note that only a single test sample was available for this class.

LIHC_T_Metastatic

The LIHC_T_Metastatic type was completely misclassified as CHOL_T_Metastatic. Note that only a single test sample was available for this class.

NET_T_Metastatic

The NET_T_Metastatic type was completely misclassified as PRAD_T_Metastatic. Note that only a single test sample was available for this class.

Cancer Type	**Precision**	**Recall**	**F1-score**	**Support**
ACC_T_Metastatic	1	1	1	1
ACC_T_Tumor	1	1	1	12
ALL_T_Metastatic	1	0.5	0.667	2
BLCA_N_Normal	1	1	1	3
BLCA_T_Metastatic	1	1	1	2
BLCA_T_Tumor	0.951	0.951	0.951	61
BRCA_N_Normal	1	1	1	17
BRCA_T_Metastatic	1	1	1	8

Cancer Type	Precision	Recall	F1-score	Support
BRCA_T_Tumor	1	1	1	165
CESC_T_Tumor	0.907	0.867	0.886	45
CHOL_N_Normal	1	1	1	1
CHOL_T_Metastatic	0.5	1	0.667	2
CHOL_T_Tumor	0.714	1	0.833	5
CLL_T_Tumor	1	1	1	4
CML_T_Tumor	1	1	1	15
COADREAD_N_Normal	1	1	1	8
COADREAD_T_Metastatic	1	1	1	1
COADREAD_T_Tumor	1	0.966	0.982	58
DLBC_T_Tumor	1	1	1	26
ESCA_T_Metastatic	0	0	0	1
ESCA_T_Tumor	0.92	0.92	0.92	25
FL_T_Tumor	0.875	1	0.933	7
GBM_T_Tumor	0.97	0.97	0.97	33
HNSC_N_Normal	0.875	1	0.933	7
HNSC_T_Metastatic	0	0	0	1
HNSC_T_Tumor	0.973	0.936	0.954	78
KICH_N_Normal	0.667	1	0.8	4
KICH_T_Tumor	0.769	1	0.87	10
KIRC_N_Normal	1	0.909	0.952	11
KIRC_T_Tumor	0.951	0.963	0.957	80
KIRP_N_Normal	1	0.8	0.889	5
KIRP_T_Tumor	0.95	0.864	0.905	44
LAML_T_Metastatic	0.5	1	0.667	1
LAML_T_Tumor	1	1	1	18
LGG_T_Tumor	0.988	0.988	0.988	80
LIHC_N_Normal	1	1	1	7

Cancer Type	Precision	Recall	F1-score	Support
LIHC_T_Metastatic	0	0	0	1
LIHC_T_Tumor	1	0.964	0.982	56
LUAD_N_Normal	0.8	0.889	0.842	9
LUAD_T_Metastatic	1	1	1	1
LUAD_T_Tumor	0.927	0.974	0.95	78
LUSC_N_Normal	0.857	0.75	0.8	8
LUSC_T_Tumor	0.932	0.92	0.926	75
MB-Adult_T_Tumor	1	1	1	41
MESO_T_Tumor	1	1	1	45
NET_T_Metastatic	0	0	0	1
OV_T_Metastatic	1	1	1	2
OV_T_Tumor	1	1	1	46
PAAD_T_Metastatic	0.5	1	0.667	1
PAAD_T_Tumor	0.964	1	0.982	27
PCPG_T_Tumor	1	1	1	28
PRAD_N_Normal	0.7	0.875	0.778	8
PRAD_T_Metastatic	0.9	1	0.947	9
PRAD_T_Tumor	0.986	0.96	0.973	75
SARC_T_Metastatic	1	1	1	5
SARC_T_Tumor	1	0.974	0.987	39
SKCM_T_Metastatic	1	1	1	2
SKCM_T_Tumor	1	0.986	0.993	71
STAD_N_Normal	1	0.8	0.889	5
STAD_T_Tumor	0.954	0.984	0.969	63
TGCT_T_Tumor	1	1	1	23
THCA_N_Normal	1	1	1	9
THCA_T_Tumor	1	1	1	77
THYM_T_Tumor	1	1	1	18

Cancer Type	Precision	Recall	F1-score	Support
UCEC_N_Normal	1	1	1	4
UCEC_T_Tumor	0.862	0.926	0.893	27
UCS_T_Tumor	0.889	0.889	0.889	9
UVM_T_Tumor	1	1	1	12
accuracy	0.965	0.965	0.965	0.965
macro avg	0.879	0.905	0.885	1723
weighted avg	0.966	0.965	0.965	1723

Table 3.6: The precision, recall, F1-score, and support for each cancer type class with testing conducted using the mixed held-out test set.

Referring to Figure 3.7, we can see that the majority of types were accurately classified with a few exceptions. The most poorly performing classes are noted above, however, as with the cancer subtype classes we again observe misclassifications within the normal lung and kidney tissue subtypes. We can also see some significant misclassifications of ALL_T_Metastatic and STAD_N_Normal.

ALL_T_Metastatic

ALL_T_Metastatic consisted of only two test samples. One of these samples was misclassified as LAML_T_Tumor.

STAD_N_Normal

STAD_N_Normal was misclassified as ESCA_T_Tumor in 20% of the samples.

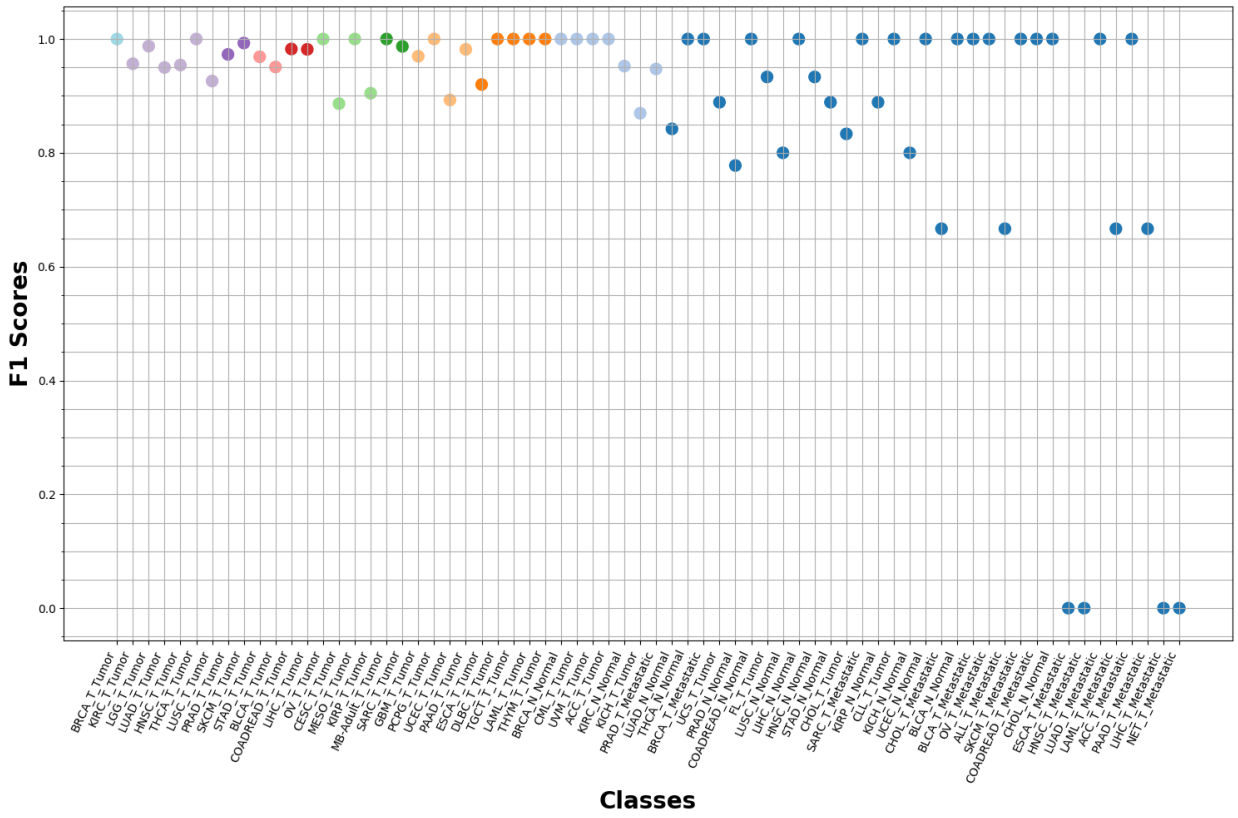

Figure 3.6: The macro F1-scores of each cancer type when testing on the held-out test set containing primary cancer, metastatic cancer, and normal tissue samples. Classes are ordered from left to right by the number of training samples available with colours representing bins of 20 samples.

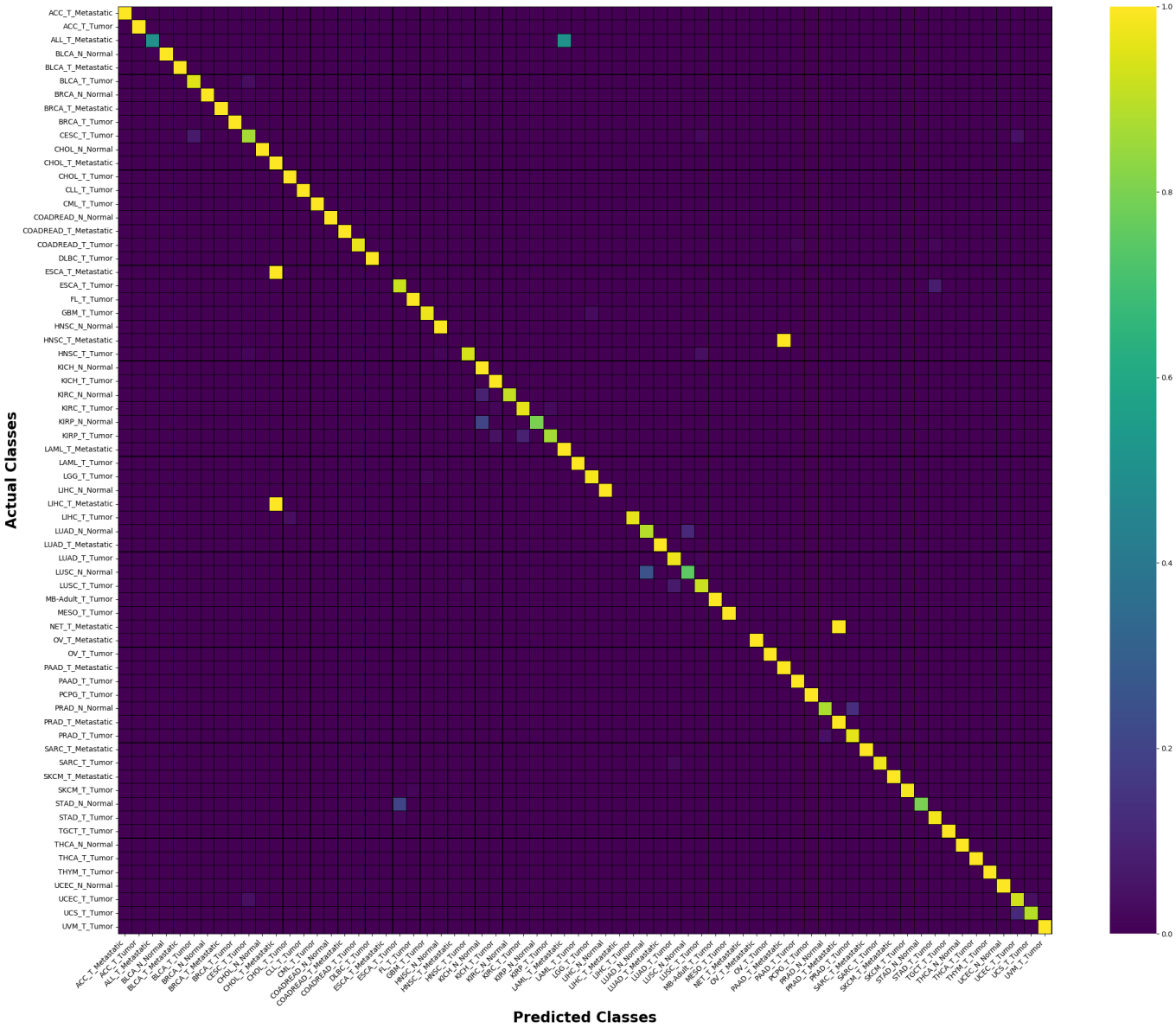

Figure 3.7: A confusion matrix depicting the cancer type classification performance on the held-out test set containing primary cancer, metastatic cancer, and normal tissue samples.

3.1.4 Cancer Subtype

The classification performance of the cancer subtype task resulted in a total F1-score of 0.885 and an accuracy of 93.3% across all 91 subtypes. The F1-score for the classification of the primary, metastatic, and normal subtypes individually was 0.851, 0.704, and 0.927

respectively (see Figure 3.5). The F1-scores, precision, and recall of each subtype can be seen in Table 3.8, and the class-wise F1-scores are presented graphically in Figure 3.8.

Cancer Subtype	**Precision**	**Recall**	**F1-score**	**Support**
ACC_T_Metastatic	1	1	1	1
ACC_T_Tumor	1	1	1	12
ALL_T_Metastatic	1	1	1	2
BLCA_N_Normal	1	1	1	3
BLCA_T_Metastatic	0.667	1	0.8	2
BLCA_T_Tumor	0.952	0.967	0.959	61
BRCA_Basal_T_Tumor	0.852	0.885	0.868	26
BRCA_HER2like_Tumor	0.818	0.75	0.783	12
BRCA_IDC_T_Metastatic	0.857	0.857	0.857	7
BRCA_ILC_T_Metastatic	0	0	0	1
BRCA_LuminalA_T_Tumor	0.819	0.951	0.88	81
BRCA_LuminalB_T_Tumor	0.724	0.677	0.7	31
BRCA_N_Normal	1	1	1	17
BRCA_T_Tumor	0.667	0.267	0.381	15
CESC_CAD_T_Tumor	1	0.571	0.727	7
CESC_SCC_T_Tumor	0.921	0.921	0.921	38
CHOL_EHCH_T_Metastatic	0	0	0	1
CHOL_IHCH_T_Metastatic	0.5	1	0.667	1
CHOL_N_Normal	1	1	1	1
CHOL_T_Tumor	0.714	1	0.833	5
CLL_T_Tumor	1	1	1	4
CML_T_Tumor	1	1	1	15
COADREAD_N_Normal	1	1	1	8
COADREAD_T_Metastatic	1	1	1	1
COADREAD_T_Tumor	1	0.983	0.991	58
DLBC_BM_T_Tumor	1	1	1	2

Cancer Subtype	Precision	Recall	F1-score	Support
DLBC_T_Tumor	1	0.958	0.979	24
ESCA_EAC_T_Metastatic	0	0	0	1
ESCA_EAC_T_Tumor	0.889	0.889	0.889	9
ESCA_SCC_T_Tumor	0.824	1	0.903	14
ESCA_T_Tumor	0	0	0	2
FL_T_Tumor	0.778	1	0.875	7
GBM_T_Tumor	1	0.97	0.985	33
HNSC_N_Normal	0.875	1	0.933	7
HNSC_T_Metastatic	0	0	0	1
HNSC_T_Tumor	0.974	0.949	0.961	78
KICH_N_Normal	0.8	1	0.889	4
KICH_T_Tumor	0.75	0.9	0.818	10
KIRC_N_Normal	1	1	1	11
KIRC_T_Tumor	0.939	0.963	0.951	80
KIRP_N_Normal	1	0.8	0.889	5
KIRP_T_Tumor	0.95	0.864	0.905	44
LAML_T_Metastatic	1	1	1	1
LAML_T_Tumor	1	1	1	18
LGG_T_Tumor	0.988	1	0.994	80
LIHC_N_Normal	1	1	1	7
LIHC_T_Metastatic	1	1	1	1
LIHC_T_Tumor	1	0.964	0.982	56
LUAD_N_Normal	0.778	0.778	0.778	9
LUAD_T_Metastatic	1	1	1	1
LUAD_T_Tumor	0.938	0.974	0.956	78
LUSC_N_Normal	0.75	0.75	0.75	8
LUSC_T_Tumor	0.958	0.92	0.939	75
MB_Group3_T_Tumor	1	1	1	6

Cancer Subtype	Precision	Recall	F1-score	Support
MB_Group4_T_Tumor	1	1	1	10
MB_SHH_T_Tumor	1	1	1	20
MB_WNT_T_Tumor	1	1	1	5
MESO_T_Tumor	1	1	1	45
OV_T_Metastatic	0.667	1	0.8	2
OV_T_Tumor	1	1	1	46
PAAD_T_Metastatic	1	1	1	1
PAAD_T_Tumor	1	1	1	27
PCPG_T_Tumor	1	1	1	28
PRAD_N_Normal	0.7	0.875	0.778	8
PRAD_T_Metastatic	0.9	1	0.947	9
PRAD_T_Tumor	0.986	0.96	0.973	75
PrNET_T_Metastatic	0	0	0	1
SARC_DDL_T_Tumor	0.857	0.667	0.75	9
SARC_LMS_T_Metastatic	1	1	1	1
SARC_LMS_T_Tumor	0.722	0.813	0.765	16
SARC_MFS_T_Tumor	0.5	0.75	0.6	4
SARC_MPNST_T_Tumor	0	0	0	1
SARC_Synovial_T_Tumor	1	1	1	1
SARC_T_Metastatic	1	1	1	4
SARC_UPS_T_Tumor	0.333	0.25	0.286	8
SKCM_T_Metastatic	1	1	1	2
SKCM_T_Tumor	1	0.986	0.993	71
STAD_CIN_T_Tumor	0.813	0.813	0.813	32
STAD_EBV_T_Tumor	0.714	1	0.833	5
STAD_GS_T_Tumor	0.692	0.818	0.75	11
STAD_MSI_T_Tumor	0.75	0.818	0.783	11
STAD_N_Normal	1	0.8	0.889	5

Cancer Subtype	**Precision**	**Recall**	**F1-score**	**Support**
STAD_T_Tumor	0	0	0	4
TGCT_T_Tumor	1	1	1	23
THCA_N_Normal	1	1	1	9
THCA_T_Tumor	1	1	1	77
THYM_T_Tumor	1	1	1	18
UCEC_N_Normal	1	1	1	4
UCEC_T_Tumor	0.893	0.926	0.909	27
UCS_T_Tumor	1	1	1	9
UVM_T_Tumor	1	1	1	12
accuracy	0.933	0.933	0.933	0.933
macro avg	0.826	0.846	0.831	1723
weighted avg	0.929	0.933	0.929	1723

Table 3.8: The precision, recall, F1-score, and support for each cancer subtype class with testing conducted using the mixed held-out test set.

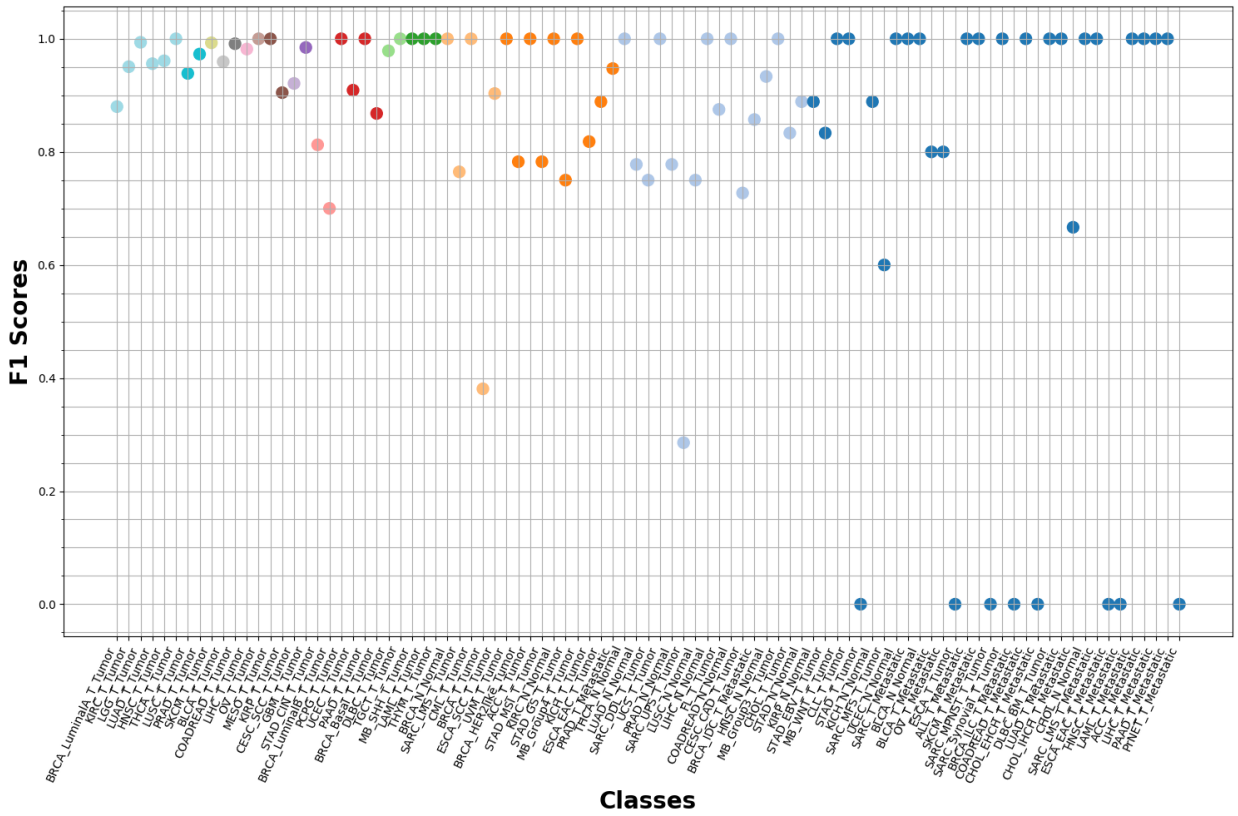

Figure 3.8: The macro F1-scores of each cancer subtype when testing on the held-out test set containing primary cancer, metastatic cancer, and normal tissue samples. Classes are ordered from left to right by the number of training samples available with colours representing bins of 20 samples.

The observed trend is that the F1-scores decrease as the number of training samples per class decreases (from left to right in Figure 3.8). The classification performance is generally better with a larger number of training samples. The largest outliers from this trend are the primary breast carcinoma (BRCA_T_Tumor) and primary undifferentiated pleomorphic sarcoma (SARC_UPS_T_Tumor) subtypes with F1-scores of 0.381 and 0.286 respectively. There are eight subtypes with F1-scores of 0.0, each of which had fewer than 20 training examples. The cancer subtype classification accuracy and the predicted classes of misclassified subtype can be seen in the confusion matrix in Figure 3.9. The eight subtypes with F1-scores of 0 and the source of their misclassifications are presented below:

STAD_T_Tumor

STAD_T_Tumor was misclassified completely with the model predicting STAD_CIN_T_Tumor, STAD_GS_T_Tumor, and STAD_MSI_T_Tumor instead. Nearly half of the STAD_T_Tumor samples were misclassified as STAD_CIN_T_Tumor.

ESCA_T_Tumor

ESCA_T_Tumor was misclassified in half the samples as ESCA_SCC_T_Tumor and the other half as STAD_EBV_T_Tumor.

SARC_MPNST_T_Tumor

SARC_MPNST_T_Tumor was misclassified completely as LUAD_T_Tumor. Note that there is only a single test sample for this subtype.

BRCA_ILC_T_Metastatic

BRCA_ILC_T_Metastatic was completely misclassified as BRCA_IDC_T_Metastatic. Note that there is only a single test sample for this subtype.

CHOL_EHCH_T_Metastatic

CHOL_EHCH_T_Metastatic was misclassified completed as CHOL_IHCH_T_Metastatic. Note that there is only a single test sample for this subtype.

ESCA_EAC_T_Metastatic

ESCA_EAC_T_Metastatic was misclassified completely as OV_T_Tumor. Note that there is only a single test sample for this subtype.

HNSC_T_Metastatic

HNSC_T_Metastatic was misclassified completely as BLCA_T_Metastatic. Note that there is only a single test sample for this subtype.

PrNET_T_Metastatic

PrNET_T_Metastatic was misclassified completely as PRAD_T_Metastatic. Note that there is only a single test sample for this subtype.

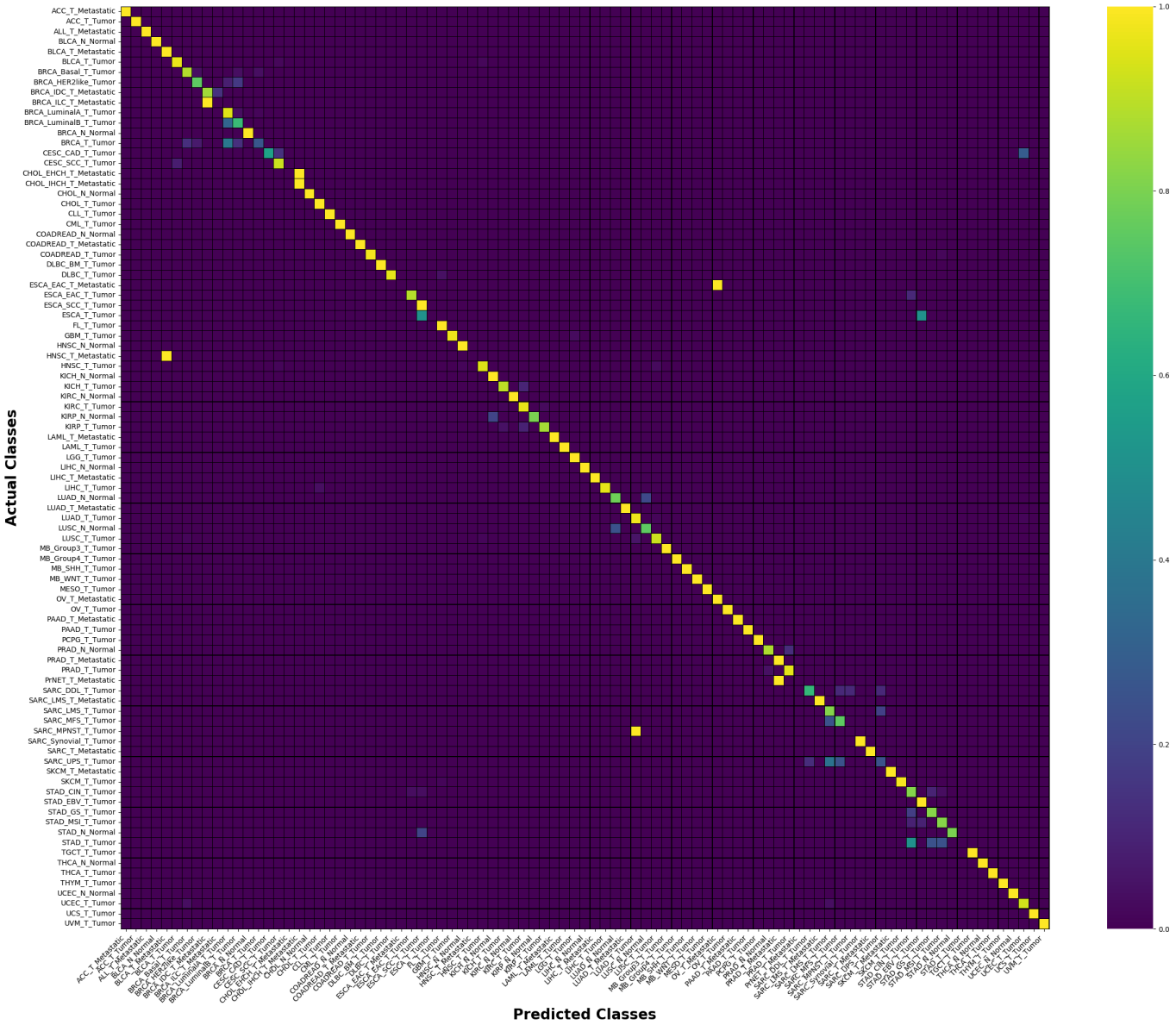

Figure 3.9: A confusion matrix depicting the cancer subtype classification performance on the held-out test set containing primary cancer, metastatic cancer, and normal tissue samples.

Referring to Figure 3.9, we can see that the majority of cancer subtypes were accurately classified, with a few exceptions. The most poorly performing classes are noted above, however, there was also misclassifications observed within the normal lung and kidney tissues and the sarcoma, stomach adenocarcinoma, and breast adenocarcinoma subtypes.

Normal Tissues

The normal tissues for the lungs and kidneys all see some misclassifications between their respective, related normal counterparts. For example, normal LUAD is misclassified as normal LUSC and normal LUSC is misclassified as normal LUAD.

STAD Subtypes

The primary stomach adenocarcinoma without a subtype annotation (STAD_T_Tumor) resulted in the largest error rate. The model completely misclassified all four test samples as either the GS, CIN, or MSI primary cancer subtypes. There were other misclassifications between STAD subtypes that can be observed in Figure 3.9. One notable observation is that the normal STAD subtype had one of five (20%) test samples incorrectly classified as ESCA_SCC_T_Tumor. This is one of two normal classes that had a misclassification as a primary cancer, with the other misclassification being PRAD (one sample called as primary PRAD).

BRCA Subtypes

The largest primary cancer offender was the primary breast cancer class without a subtype annotation (BRCA_T_Tumor). It consisted of 15 test samples and was misclassified 75% of the time as a mixture of the other primary cancer subtypes. As mentioned above, BRCA_ILC_T_Metastatic was classified incorrectly as BRCA_IDC_T_Tumor for every test sample.

SARC Subtypes

Sarcoma subtypes had misclassifications observed within the primary LMS, DDL, MFS, and MPNST subtypes. As mentioned above MPNST was completely misclassified. The UPS

subtype was the only subtype in which the majority of its samples were classified incorrectly. All of the other subtypes had fewer than half of their samples misclassified.

ESCA Subtypes

The primary ESCA subtypes generally had much better classification performance than the BRCA, STAD, and SARC subtypes. The poorest performer was again the class without a subtype annotation. Primary ESCA_EAC had a single sample misclassified as STAD_CIN and the Metastatic ESCA_EAC was completely misclassifed (as mentioned above).

MB Subtypes

The medulloblastoma subtypes were all classified with perfect accuracy and F1-scores.

3.2 Discussion: Held-out Test Set Classification

All of the misclassifications that occurred within the held-out test set remained within the same disease type. We do not see any primary cancer samples classified as metastatic cancers and vice versa. Even when a sample is classified to a subtype in a completely different organ system, like with the DLBC_T_Metastatic subtype, it is misclassified as another metastatic cancer. This implies that model has learned to distinguish differences in the expression patterns of metastatic samples when compared to primary one. We will see, however, that this does not hold true for samples from the POG test set.

3.2.1 Normal Tissue

Cross-calling was observed within the normal lung and kidney classes in both the cancer type and subtype tasks. Cross-calling in this context refers to two or more classes that

are misclassified as each other in at least some of the samples. This cross-calling was expected behaviour within the normal classes and serves as a sanity check of sorts. The normal kidney and lung classes are labelled based on the adjacent tumour type/subtype. For example, a patient with a primary LUSC tumour (LUSC_T_Tumor) would have their tumour adjacent normal lung sample labelled as LUSC_N_Normal. A patient with primary LUAD (LUAD_T_Tumor) would have their lung normal sample labelled as LUAD_N_Normal. So, while these patients' normal lung samples have different labels, they are both normal lung tissue and could have been given identical class labels. Keeping these labels separate by type/subtype allows us to confirm that the model is indeed learning the underlying biology of the samples. Seeing cross-calling between normals of the same tissue (ie. lung or kidney) indicates that the model has learned what characterizes these tissues and is interchanging their labels as a result.

3.2.2 Complete Misclassifications

The misclassifications of the largest concern are those that were completely misclassified and have no apparent biological underpinning (see Sections 3.1.4 and 3.1.3). We saw examples of these using both the held-out and external test sets on the cancer type and cancer subtype learning tasks. In the held-out test set on the cancer type task, we saw four examples of total misclassifications. Each of these examples were metastatic cancer types and all only consisted of a single sample. The small sample sizes make these results potential aberrations and are not conclusive enough to be considered total faults of the learning task. At the cancer subtype task, however, we had more test samples with which to conclude the poor performance of the model. We did again see metastatic subtypes that contained only single samples, so I will exclude these from further discussion. I will also exclude the SARC_MPNST_T_Tumor subtype as there was also only a single test sample for this class.

The two remaining multi-sample complete misclassifications (STAD_T_Tumor and ESCA_T_Tumor) were for subtypes that did not have a subtype annotation. It is not entirely clear what sub-

type would be the correct annotation for these samples as TCGA subtyping excludes some samples in some cases and in others does not assign a subtype based on not matching entirely with the subtype annotations they defined [**66**, **67**, **68**, **69**]. Further in depth analysis of each sample and the TCGA subtyping protocols may reveal further information pertaining to these samples and may help to clarify the model's performance on them.

3.2.3 Cancer Type and Subtype Performance Comparison by Disease State

Figure 3.5 illustrates that primary cancers were better classified within the cancer type task than within the cancer subtype task. The metastatic cancers were classified better within the cancer subtype task, and the normal samples were classified with similar performance in both tasks. It is important to note that there are very few metastatic subtypes when compared to primary subtypes. As a result, the class sizes decrease more within the primary cancers when moving from cancer type to subtype and may have contributed to the reduced classification performance seen for the primary cancers within the subtype task.

3.3 Results: Metastatic-Only External (POG) Test Set

3.3.1 Organ System of Origin

The classification F1-scores of the model can be seen in Table 3.2 and Figure 3.10. A confusion matrix depicting the rate and subject of classification errors can be seen in Figure 3.11.

Overall, the performance of the organ system classifications on the POG test set resulted in an F1-score that is 0.095 less than we saw with the held-out test set. The F1-scores range from 0.6 to 1.0 (Figure 3.10. The poorest performing class is head and neck, and

the two best (F1-scores of 1.0) performing classes are endocrine and hematologic. The head and neck class was misclassified approximately 20% of the time as either gynecologic or soft tissue (Figure 3.11).

Organ System of Origin	**Precision**	**Recall**	**F1-score**	**Support**
Breast	0.964	1	0.982	134
Endocrine	1	1	1	6
Gastrointestinal	0.988	1	0.994	163
Gynecologic	0.962	0.735	0.833	34
Head and Neck	0.6	0.6	0.6	5
Hematologic	1	1	1	2
Skin	0.933	1	0.966	14
Soft Tissue	0.959	0.839	0.895	56
Thoracic	0.889	0.909	0.899	44
Urologic	0.6	1	0.75	3
macro avg	0.889	0.908	0.892	461
weighted avg	0.958	0.948	0.951	461

Table 3.10: The precision, recall, F1-score, and support for each organ system of origin class with testing conducted using the metastatic-only external (POG) test set.

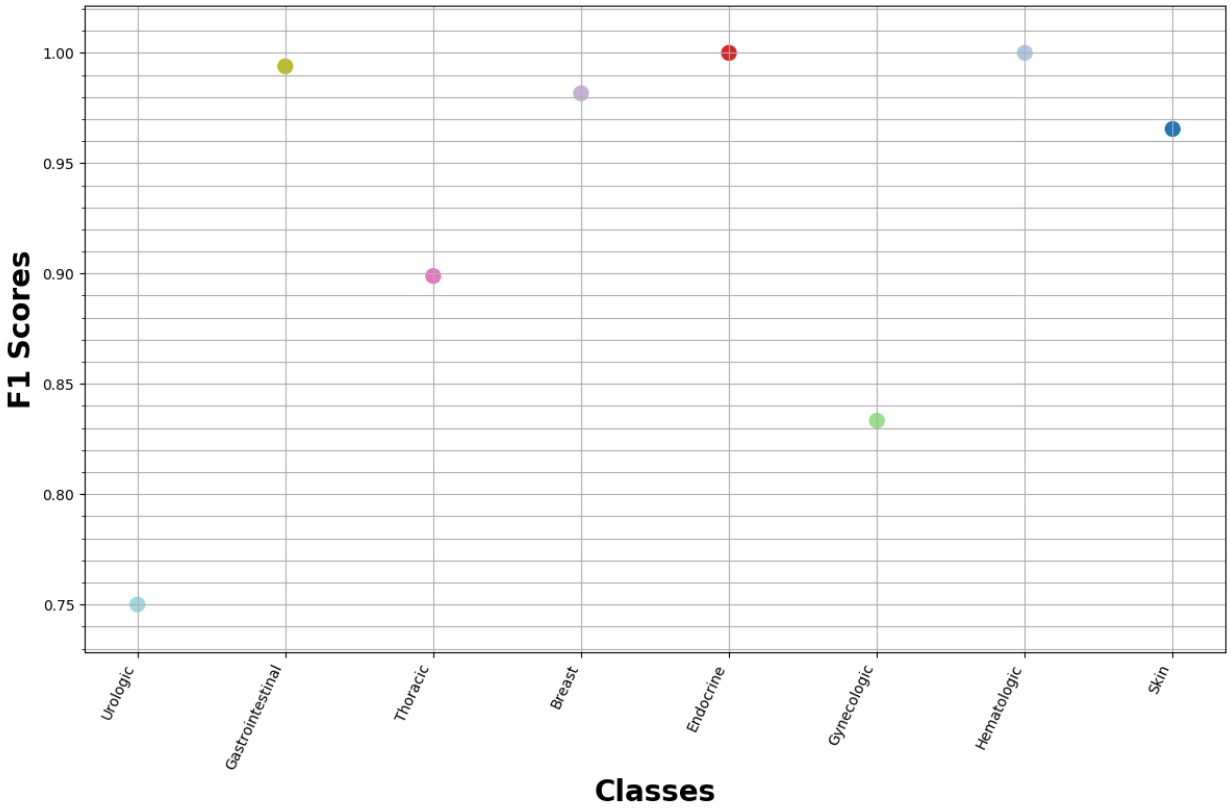

Figure 3.10: The macro F1-scores of each organ system of origin when testing on the metastatic-only external (POG) test set. Classes are ordered from left to right by the number of training samples available with colours representing bins of 20 samples.

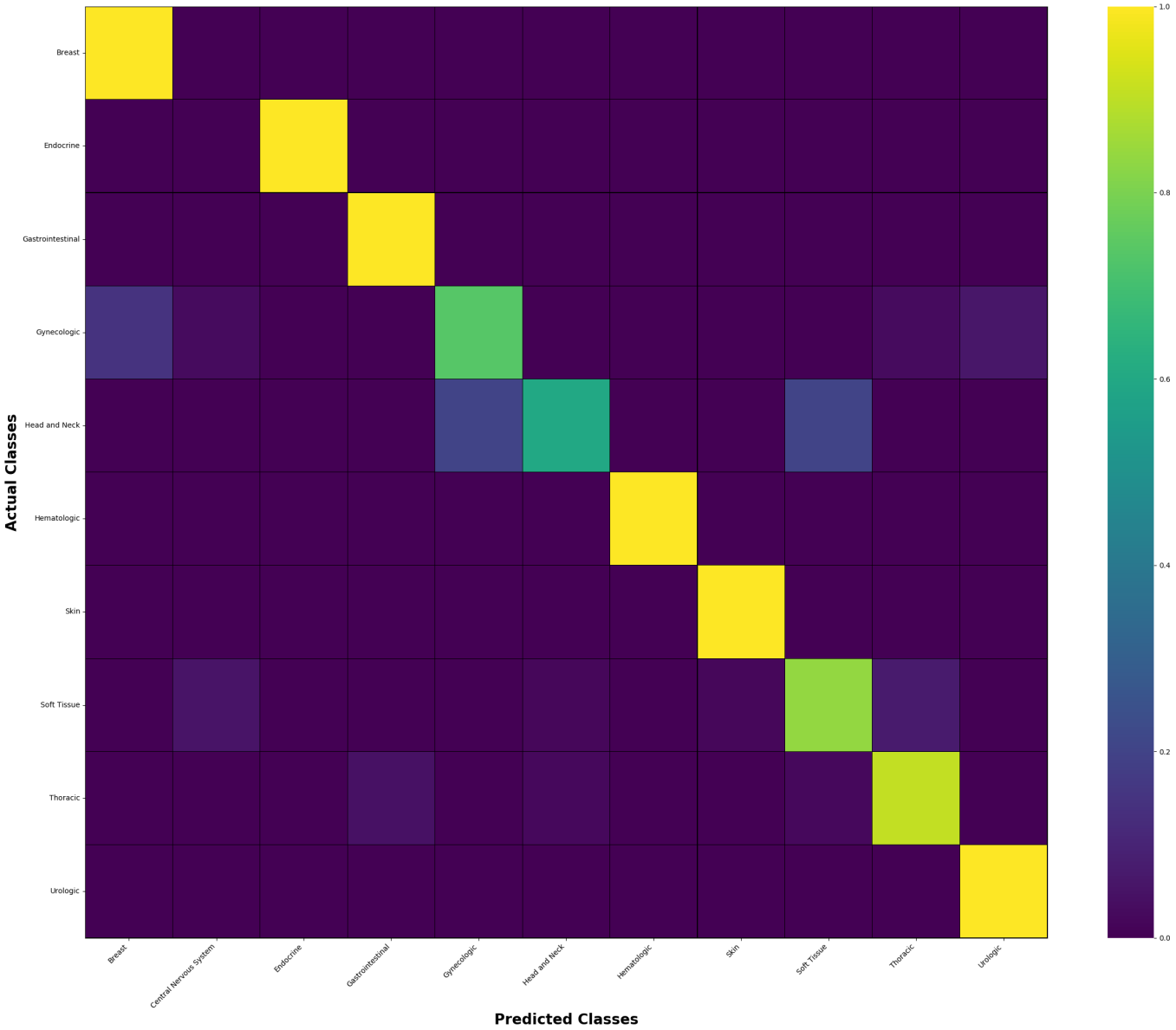

Figure 3.11: A confusion matrix depicting the organ system of origin classification performance on the metastatic-only external (POG) test set.

3.3.2 Disease State

The classification performance of this learning task was the greatest of the four tasks. Out of the 495 test samples, 461 were correctly classified correctly and resulted in an F1-score of 0.886 (see Table 2.8. This is a decrease in the F1-score of 0.114 when compared to the

metastatic classification performance with the mixed held-out test set.

Disease State	Precision	Recall	F1-score	Support
Metastatic	1	0.796	0.886	461
macro avg	1	0.796	0.886	461
weighted avg	1	0.796	0.886	461

Table 3.12: The precision, recall, F1-score, and support for each disease state class with testing conducted using the metastatic-only external (POG) test set.

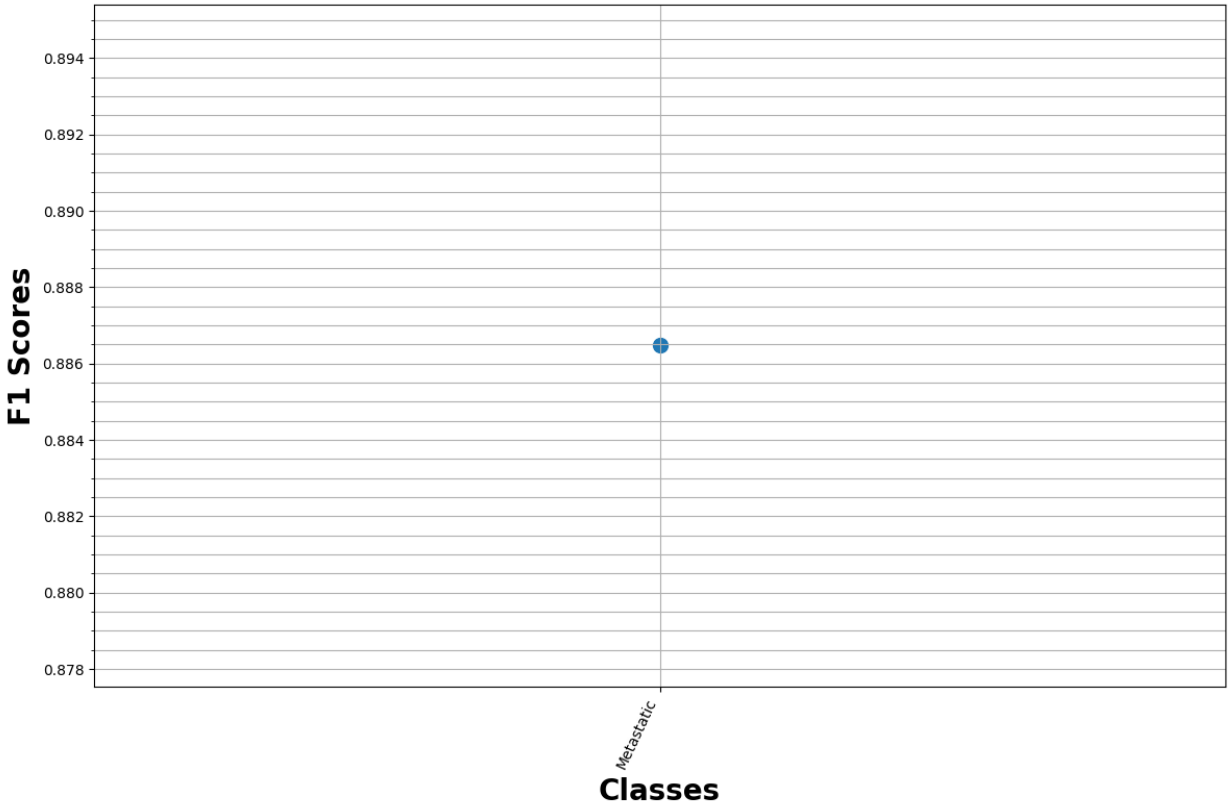

Figure 3.12: The macro F1-scores of each disease state when testing on the metastatic-only external (POG) test set. Classes are ordered from left to right by the number of training samples available with colours representing bins of 20 samples.

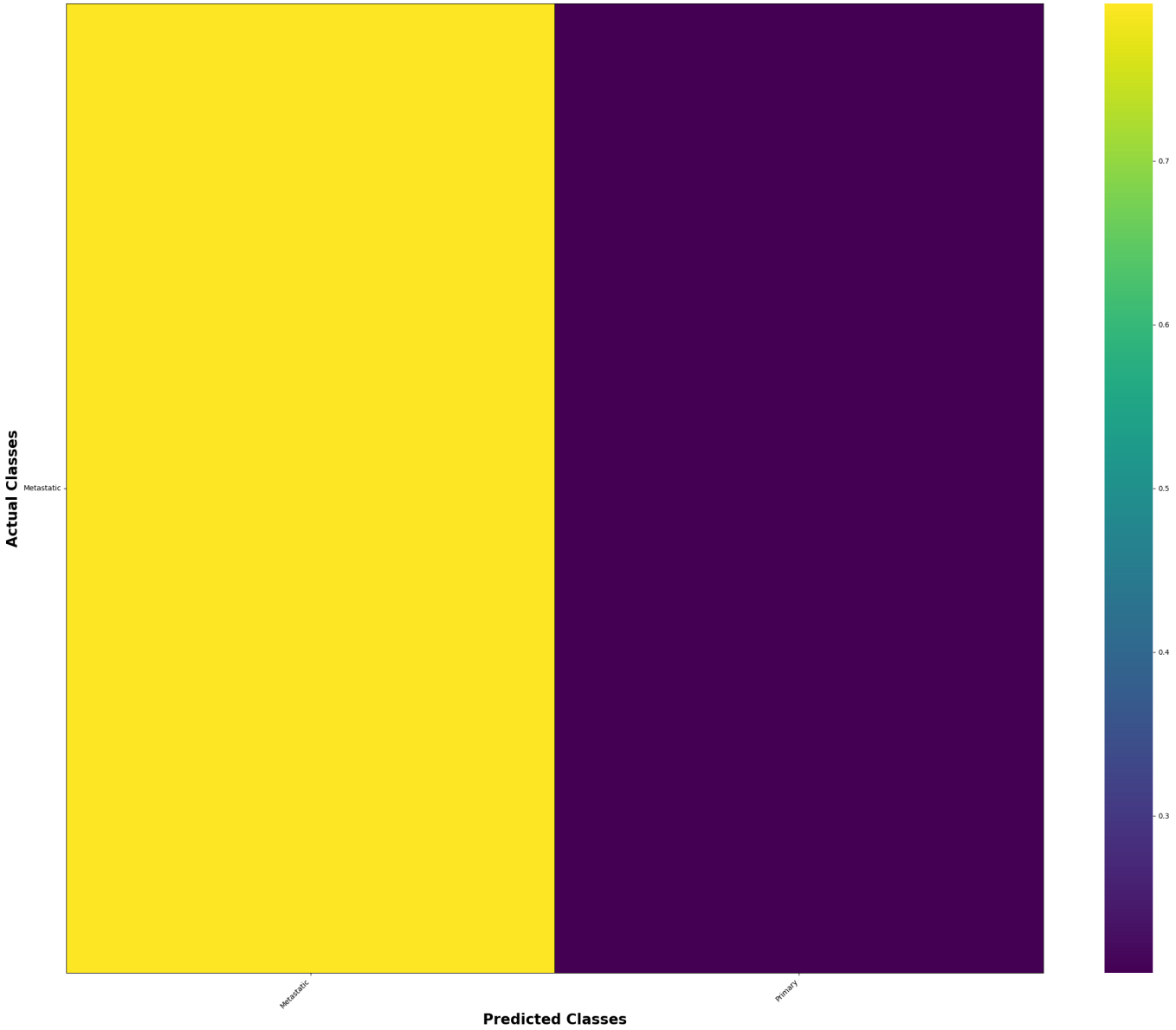

Figure 3.13: A confusion matrix depicting the disease state classification performance on the metastatic-only external (POG) test set.

3.3.3 Cancer Type

The classification performance of the models on the metastatic-only test set resulted in an overall F1-score of 0.761 (Table 3.6. The F1-scores, precision, and recall of each class can be seen in Table 3.16 and the F1-scores are presented visually in Figure 3.16. A confusion

matrix presenting the predicted classes can be seen in Figure 3.15. Overall, the classification performance on the cancer type learning task using the metastatic-only test set resulted in an F1-score that was lower by 0.124 when compared to the results obtained using the held-out set. We also observe no easily observed downward trend of classification performance as a function of training class size like we saw with the held-out test set (Figures 3.14 and 3.14).

The F1-scores ranged from 0.0 to 1.0 with LAML_T_Metastatic being completely misclassified and the ACC_T_Metastatic and PRAD_T_Metastatic classes being completely accurately classified. Referring to Figure 3.15, we observe that seven primary tumour and one normal class were incorrectly predicted. The LUAD_T_Metastatic class is observed to have the largest number of incorrectly predicted classes, though still maintained an F1-score of 0.725.

We can also observe a large portion of gastric cancers being misclassified as another gastric cancer. PAAD_T_Metastatic samples were predicted with high frequency as CHOL_T_Metastatic or ESCA_T_Metastatic and COADREAD_T_Metastatic was frequently predicted incorrectly as ESCA_T_Metastatic.

Cancer Type	**Precision**	**Recall**	**F1-score**	**Support**
ACC_T_Metastatic	1	1	1	6
BRCA_T_Metastatic	0.978	0.993	0.985	134
CHOL_T_Metastatic	0.176	1	0.3	3
COADREAD_T_Metastatic	0.957	0.788	0.865	85
HNSC_T_Metastatic	0.8	0.8	0.8	5
LAML_T_Metastatic	0	0	0	2
LIHC_T_Metastatic	0.75	1	0.857	3
LUAD_T_Metastatic	1	0.568	0.725	44
OV_T_Metastatic	1	0.824	0.903	34
PAAD_T_Metastatic	0.822	0.514	0.632	72

Cancer Type	Precision	Recall	F1-score	Support
PRAD_T_Metastatic	1	1	1	3
SARC_T_Metastatic	0.94	0.839	0.887	56
SKCM_T_Metastatic	0.875	1	0.933	14
macro avg	0.792	0.794	0.761	461
weighted avg	0.933	0.803	0.852	461

Table 3.14: The precision, recall, F1-score, and support for each cancer type class with testing conducted using the metastatic-only external (POG) test set.

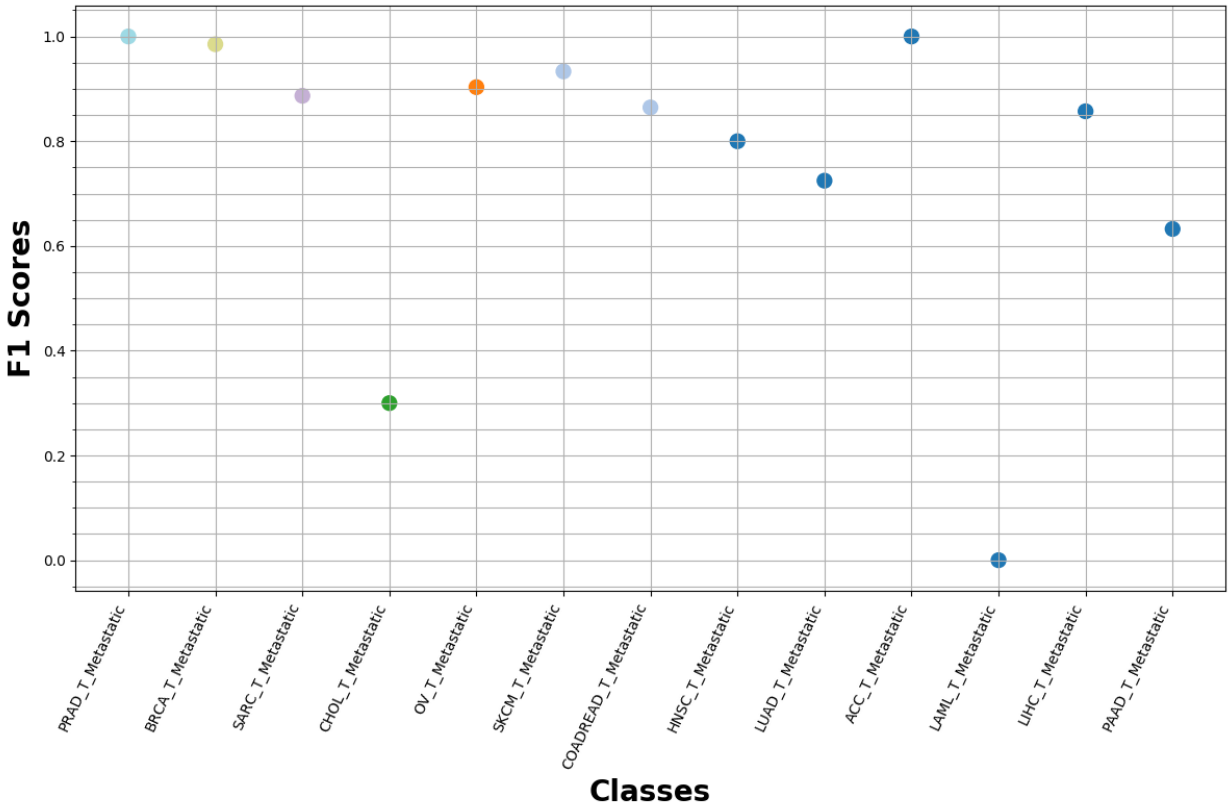

Figure 3.14: The macro F1-scores of each cancer type when testing on the metastatic-only external (POG) test set. Classes are ordered from left to right by the number of training samples available with colours representing bins of 20 samples.

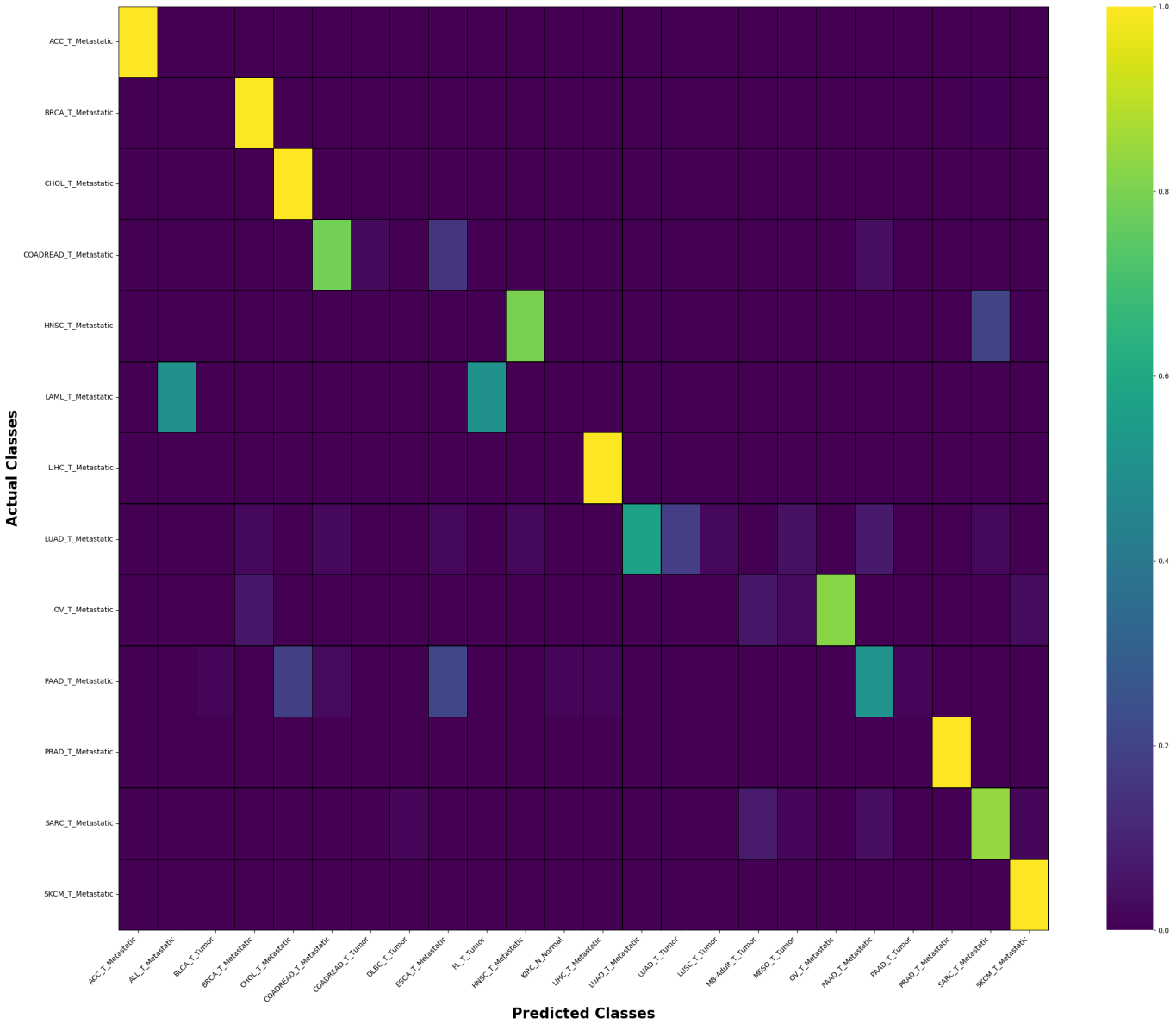

Figure 3.15: A confusion matrix depicting the cancer type classification performance on the metastatic-only external (POG) test set.

3.3.4 Cancer Subtype

The classification performance of the models on the metastatic only data set resulted in an F1-score of 0.683 and an accuracy of 67.0%. The F1-scores, precision, and recall of each class can be seen in Table 3.16 and the F1-scores are presented visually in Figure 3.16.

Cancer Subtype	Precision	Recall	F1-score	Support
ACC_T_Metastatic	1	1	1	6
BRCA_IDC_T_Metastatic	0.897	0.696	0.784	125
BRCA_ILC_T_Metastatic	0	0	0	9
CHOL_IHCH_T_Metastatic	0.158	1	0.273	3
COADREAD_T_Metastatic	0.941	0.753	0.837	85
HNSC_T_Metastatic	0.667	0.8	0.727	5
LAML_T_Metastatic	0	0	0	2
LIHC_T_Metastatic	1	1	1	3
LUAD_T_Metastatic	1	0.636	0.778	44
OV_T_Metastatic	0.966	0.824	0.889	34
PAAD_T_Metastatic	0.889	0.444	0.593	72
PRAD_T_Metastatic	1	1	1	3
SARC_LMS_T_Metastatic	0.643	0.818	0.72	11
SARC_T_Metastatic	0.824	0.622	0.709	45
SKCM_T_Metastatic	0.875	1	0.933	14
macro avg	0.724	0.706	0.683	461
weighted avg	0.879	0.67	0.75	461

Table 3.16: The precision, recall, F1-score, and support for each disease state class with testing conducted using the metastatic-only external (POG) test set.

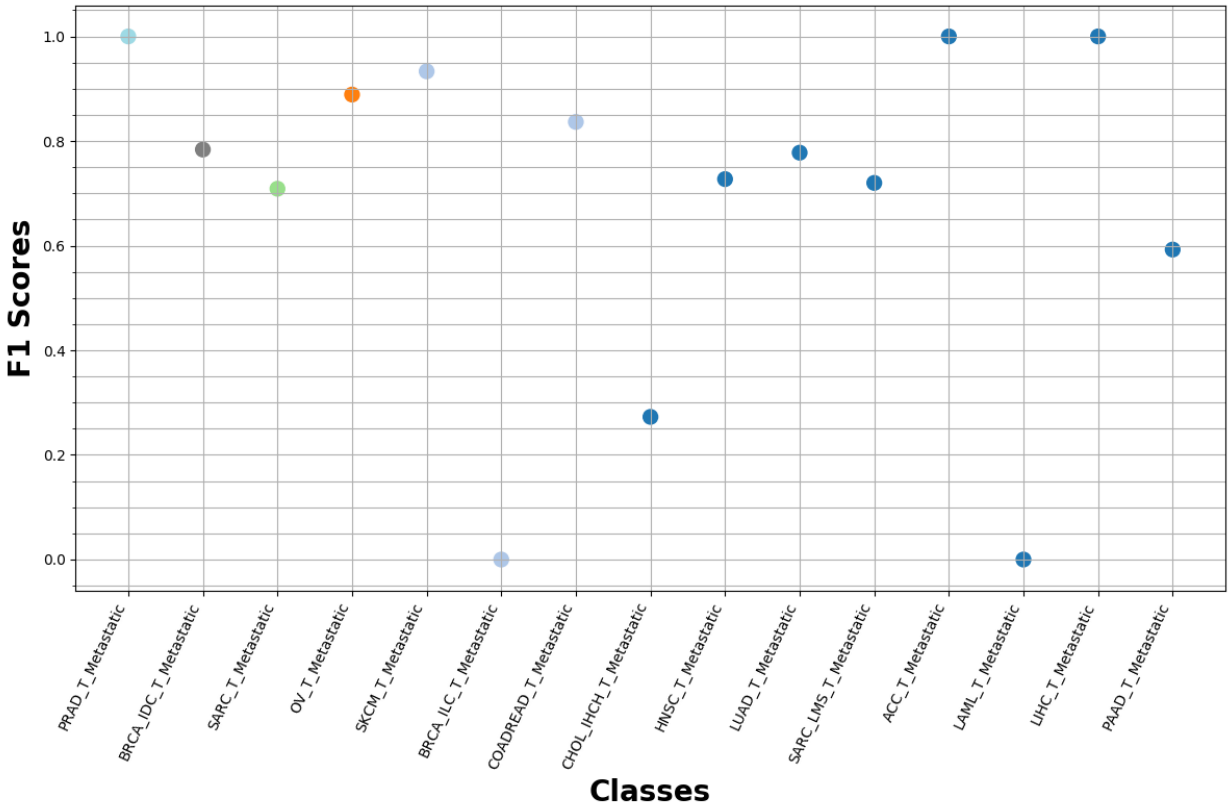

Figure 3.16: The macro F1-scores of each cancer subtype when testing on the metastatic-only external (POG) test set. Classes are ordered from left to right by the number of training samples available with colours representing bins of 20 samples.

In contrast to the held-out test set results, there appears to be no downward trend between the number of training samples and classification performance. However, eight of the 15 classes have fewer than 20 training samples with two of three of the worst performing classes being in this category (depicted in dark blue in Figure 3.16). The majority of classes (12 of 15 subtypes) achieved F1-scores between 0.593 and 1.0, with the exception BRCA_ILC_T_Tumor, LAML_T_Metastatic, and CHOL_IHCH_T_Metastatic which obtained F1-scores of 0.0, 0.0, and 0.273 respectively. However, there were two subtypes that had perfect classification accuracy: ACC_T_Metastatic, LIHC_T_Metastatic, and PRAD_T_Metastatic.

Looking at the predicted classes in Figure 3.17, we see a number of misclassifications that span disease types. There are 15 primary tumour subtypes called and one normal. The largest offending class for misclassifying samples as primary cancer subtypes was the SARC_T_Metastatic class with 5 primary subtypes called.

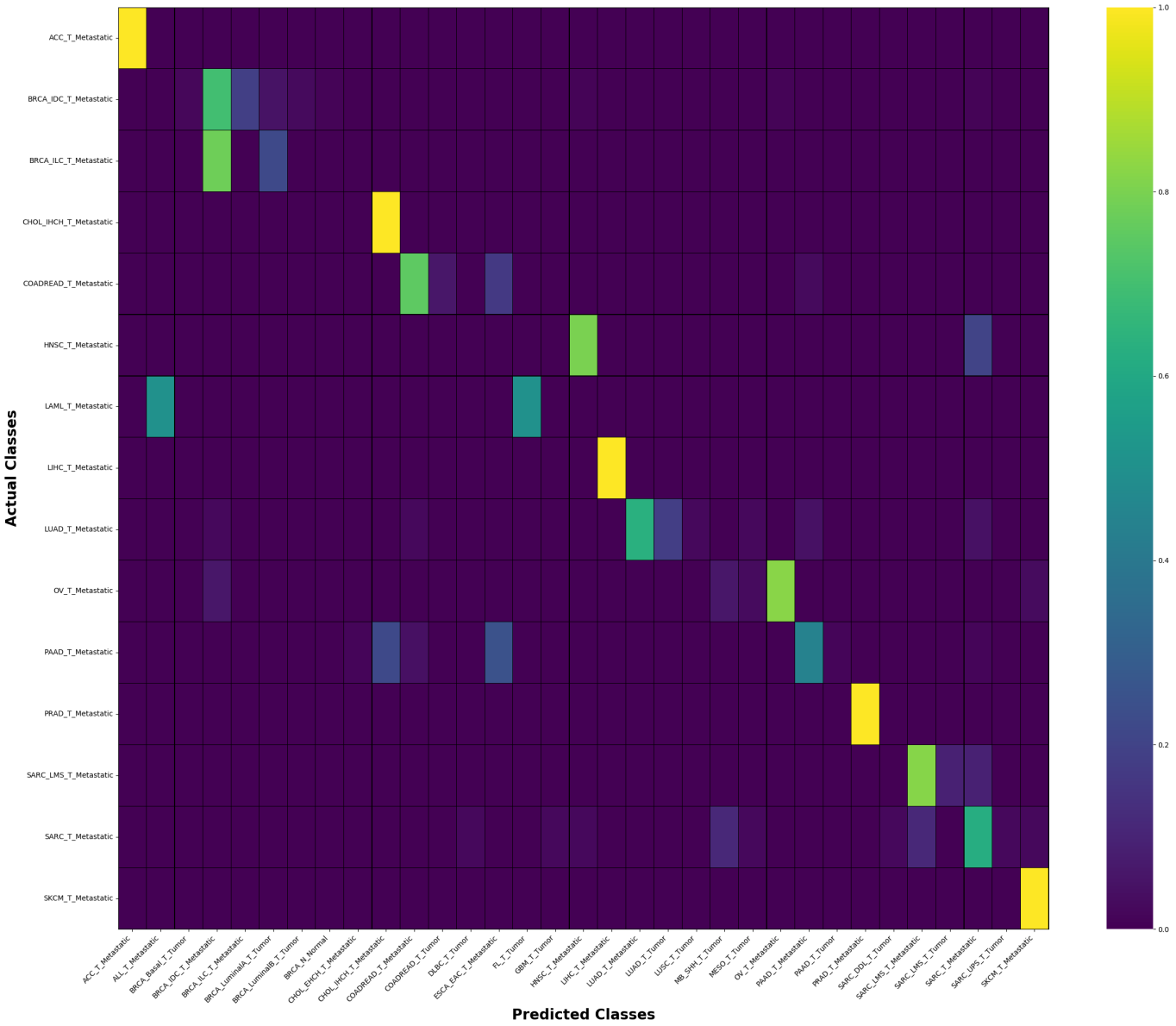

Figure 3.17: A confusion matrix depicting the disease state classification performance on the metastatic-only external (POG) test set.

Figure 3.17 also shows a much higher incidence of gastric cancers being classified as other gastric cancers than we observed in the held-out test set. We can see COADREAD_T_Metastatic was called as ESCA_T_Tumor and PAAD_T_Metastatic called as CHOL_IHCH_T_Metastatic, CHOL_EHCH_T_Metastatic, ESCA_EAC_T_Metastatic, and PAAD_T_Tumor cancer.

Generally, the overall classification performance observed with the external, metastatic-only test set is worse than the performance seen with the mixed held-out test set. The spread of subtype misclassications also generally spans more incorrect subtypes with greater frequency and variety than the misclassifications within the held-out set.

3.4 Discussion: POG Test Set Classification

Aside from the decrease in classification performance using the metastatic-only test set when compared to the held-out test set, there are two other aberrations. The first is the number of predictions of classes outside the same disease type as the true class. The second aberration is that there is a larger range of incorrect classes predicted per test class than testing on the held-out set. On the POG test set, the cancer subtype learning task had 15 primary tumour classes predicted and one normal, whereas on the held-out test set we saw no classifications of metastatic cancers as primary or normal classes. To the second point, the metastatic-only test set had classes that predicted 9 or 10 different classes (SARC_T_Metastatic and BRCA_IDC_T_Metastatic) whereas the held-out set only had at most 2 predicted classes within the metastatic subtypes (BRCA_IDC_T_Metastatic). These two aberrations are significant because they indicate the uncertainty with some samples was so great that it overcame the information learned by the model at upstream learning tasks. For example, we can observe in Figure 3.17 that approximately one quarter of the metastatic LUAD samples were misclassified as primary LUAD. This same classification error is repeated at the cancer type level. However, the amount of misclassification seen within the disease state task does not seem to indicate the same level of confusion as fewer than 7% of all samples were classified as primary cancer. This implies that the model was able

to correctly identify many samples as metastatic, but that in downstream learning tasks, the level of uncertainty was high enough to outweigh the information passed from upstream layers. This is an effect not seen using the held-out test set even within the metastatic samples.

What these aberrations imply is that there is something about the external test set that is making it more difficult for the model to make accurate predictions. A potential cause for this is that the external test set data was derived at a different facility than the data used in training. There are differences in the preparation and sequencing process that may have produced variations in the output [**60**, **62**]. It is possible the model is suffering the negative consequences of batch effect. There are two ways to potentially mitigate this effect in the context of metastatic samples. The first would be to include a larger variety of training samples from different facilities and sequencing protocols and the second would be to apply batch correction to the data prior to training. Essentially the model is, at least to some extent, fitting to noise in the data produced by the sequencing process (ie. batch effect).

3.5 Discussion: Summary

As the granularity of the classification task increases with respect to biological complexity, we do see a coinciding decrease in classification performance of the model. As such, the cancer subtype learning task resulted in the poorest classification performance. In addition to the increased complexity of this learning task relative to the other tasks, the model is also faced with smaller class sizes for each subtype. The class size reduction is a result of increasing label granularity while maintaining the size of the training data. The combination of these two factors produce smaller classes and thus reduce the training data size for the affected classes. In general, with all things being equal and over-fitting being carefully managed, a smaller training data set will always result in poorer performance when compared to a larger data set. We can speculate that having access to more subtype data could mitigate the effect of increased classification task complexity.

The ability of the model to learn each task and each class was not uniform. As we saw above, some classes, particularly within the cancer type and subtype tasks, were classified with complete accuracy, while others were entirely misclassified. One contributing factor to this outcome is the class distribution. The training data ranges from classes of size 6 (Primary PrNET) to classes of size 532 (Primary KIRC). The impact of class distribution is well-studied in machine learning and there are techniques to overcome this [**70**, **71**, **72**]. However, these are often not able to fully mitigate the effect of class imbalance and in the domain of thesis, was further impacted by the high-dimensionality of the data. High-dimensionality in machine learning is defined as having a much larger set of input features than the number of input samples for a given data set [**73**]. The effect of high-dimensionality as it relates to machine learning often hinges on the complexity of the model. Complex models, like large neural networks, that perform well in a high-dimensional domain quite often are in fact over-fitting to certain features of the data and suffer a reduced ability to generalize during prediction when compared with simpler models [**73**, **74**]. By coupling high-dimensionality and large class imbalances together, we have two factors that play a significant part in the poor performance of the model on learning the minority classes. Further experimentation with techniques that address class imbalance and feature selection may provide improved performance for this model going forward. If feature selection is considered, however, we would have to carefully consider the impact of this on post-classification analysis.

Within the held-out test set results, the majority of subtypes had the bulk of their samples correctly classified. With the exception of medulloblastomas, the model shows room for improvement among subtype classifications. We saw particularly poor performance in the cancer subtype task with the classes that lacked subtype annotations. The TCGA data set excludes subtype annotation for samples for a number of reasons including sample duplication, unknown subject identity, low DNA/RNA yield, unacceptable histology, or failed pathology review among others [**67**]. This criteria is not fully explained and may differ between cancer types. The ambiguous nature of the underlying biology of these classes, coupled with their mixed classifications, suggests that the model may be improved by excluding these classes in future iterations of training in order to better train the model to recognize

subtypes. By excluding these classes in the future, we would remove one source of ambiguity for the model as well as improve our ability to analyse the model's true subtype learning performance.

The classification performance of the model is generally satisfactory across all learning tasks in both mixed and metastatic-only domains. There is absolutely room for improvement, but the classification performance is higher than we would expect to see using random class assignment. Metastatic cancers being the form of cancer suffering from the highest misdiagnosis rates and poorest classification performance using the neural network model in this thesis, this is a good metric to determine the model's efficacy compared to current diagnostic practices. Studies have indicated that the misdiagnosis rates of metastatic cancers using standard pathological analysis can range from 45% to 94% [**61**, **74**]. The model used in this thesis has shown performance that improves upon this misdiagnosis rate and achieves an F1-score of 0.724 at its poorest performing classification task (cancer subtype using the external test set). Furthermore, according to a 2010 literature review by Anderson and Weiss, correct tissue identification from metastatic-only samples was 65.6% using immunohistochemical analysis [**75**]. Again, the performance of the neural network model exceeds this achieving an F1-score of 1.0 on a similar diagnostic task (organ system of origin classification from metastatic-only samples).

The model's overall classification performance indicates that we can, with some confidence, state that it has learned to represent the biology of and predict the correct class for a variety of cancers. With sufficient performance established, it is viable to utilize the trained model to extract and analyze the important features learned by the model in search of biological insights. This is the focus of the remainder of this thesis beginning at Chapter 4.

Chapter 4

Deeplift Analysis

4.1 Methods

4.1.1 DeepLift

After creating the fully-trained multi-task neural network model described in the previous chapters, DeepLift was run on the model to obtain a sample-gene importance score matrix. DeepLift was run sample-wise on each classification task separately. The samples provided to DeepLift were simply all of the samples in the training set used to train the model. This data is described in Section 2.1. Since DeepLift computes scores for each gene on each sample on each class within a single classification task, the resulting output is n matrices for each classification task where n is the number of classes within that classification task. For clarity, there were 91 matrices created for the cancer subtype task because there are 91 cancer subtypes as possible output classifications of the model. Each matrix is of size $s * g$ where s is the number of samples and g is the number of genes. In this case, $s = 9736$ and $g = 26668$ resulting in matrices of size $9736 * 26668$. In total, DeepLift produced 3 sets of $9736 * 26668$ matrices for the disease state task, 11 for the organ system of origin, 68 for the cancer type, and 91 for the cancer subtype task.

Due to the stochastic nature of neural network training (SGD and weight initialization), as well as the complexity of both the problem domain and the models themselves, training multiple iterations of the same model with the same data would most likely produce differing solutions for the classification tasks outlined in this thesis. For this reason, to further improve the robustness of the results from DeepLift, five models were trained and DeepLift was run on each. Each model was trained using the same set of training data and hyperparameters as outlined in Section 2.2, and DeepLift was run on each model in the fashion described above. This resulted in five sets (one for each model) of matrices corresponding to each classification task with each matrix containing per-sample importance scores. These scores were then averaged within each matrix across all samples to obtain five sets of matrices with single average values within them. Each averaged matrix was then averaged across all five models. The final result being a single set of importance scores for each gene for each of the classes across each classification task that is derived from across five models. The final format of the data is 11, 3, 68, and 91 vectors of length 26668 corresponding to the organ system of origin, disease type, cancer type, and cancer subtype learning task classes. In other words, we have importance scores for each gene that reflects how important that gene is for a positive classification of each class across all of the classification tasks.

4.1.2 Interpreting Gene Lists

For the following results and discussion, the list of the important genes for each class excludes genes with importance scores at or below 0. This means that for each class, we are only considering the genes that were influential in making a positive classification of the class in question.

In order to gain insight into the functionality of gene lists, annotation of functionally enriched pathways was conducted using DAVID [**76**, **77**]. The default settings were used in all cases. When considering enrichment scores in the results presented below, pathways or clusters were considered significant if they had an enrichment score > 1.3, as that is

equivalent to a p-value < 0.05 [**76**].

4.1.3 Over and Underexpression Calculation

In the subsections to follow, some results are presented on the over and underexpression of genes. A gene was considered overexpressed when its mean RPKM value within the class in question was greater than two standard deviations above the mean RPKM value across all other classes within the same classification task. Similarly, a gene was considered underexpressed if its mean RPKM value was lower than two standard deviations below the mean across all other classes in the same task. The mean and standard deviation were calculated using the *pandas* Python package *mean* and *std* functions [**63**].

4.2 Results and Discussion

The results presented here were taken from the average, positively scored genes (identified by DeepLift) as described in the preceding section. It is important to note that the scale of the data available for analysis is large and the analysis conducted here is far from exhaustive. This section will focus primarily on larger trends with some examples within specific classes. In addition to this, the results are divided into subsections and are interleaved with their relevant discussion for better readability.

The results have been examined in five different ways. Each of these ways exemplifies one aspect of the data that can be challenged and studied in further depth in future work. The first is examining how many important genes were identified for each class. The second is looking for any patterns of over or underexpression of genes within the identified important genes. The third is to examine the functionality of highly enriched pathways within different classes. The fourth and fifth analyses involve quantifying the role of RNA genes and pseudogenes within the important genes for various tasks and classes.

4.2.1 Validation of Results: Normal Tissues

It is important to ensure that the model is indeed finding genes that accurately reflect the biology of the included classes. One way we can do this is to leverage the normal tissue classes embedded in each classification. We can examine the important genes for classifying normal breast tissue, for example, and look for genes involved in lactation. By doing this for a few tissues with specialized functions, we can offer some validity to the results as they pertain to the cancer classes and further increase our confidence in the DeepLift results. To obtain the functional annotations, the DAVID functional annotation tool was used and the functional annotation chart was examined for relevant pathways [**76**].

It is important to note that the secreted, signal, signal peptide, and extracellular region pathways are highly enriched in the normal tissue classes as identified using DAVID on the disease state important genes for the normal class. For the results to follow I will exclude discussing these pathways as being significant because they are not tissue specific and are likely used by the model to distinguish normal classes from cancer classes.

Thyroid

The first normal class examined was normal thyroid tissue: THCA. Since the thyroid performs a small set of very specific functions, the genes used to identify it should, at least in part, reflect this functionality. The important genes from the cancer type classification task were used to validate if the model is identifying genes of biological relevance. The cancer type class was selected because its the first task that forces the model to identify thyroid tissue explicitly. In theory, it is here that the model will need to distinguish thyroid genes. The functions for the top 10 (ordered by descending p-value) enriched pathways can be seen in Figure 4.1. DAVID identified 4 out of the 123 important genes identified by DeepLift as being a part of a thyroid hormone generation pathway with a highly significant p-value of 3.8E-5. There are two other notable indications of thyroid tissue. The first is the neuropeptide hormone activity pathway. A number of neuropeptides are found within the thyroid and thus

can be indicative of thyroid tissue [**78**]. The second is the olfactory transduction pathway. A cursory search through the genes involved in this pathway (according to KEGG) were shown to be highly expressed in the thyroid by the Genotype-Tissue Expression (GTEx) Project [**79**, **80**, **81**]. The significance of these pathways according to DAVID is indicative of the identified genes having biological relevance to thyroid tissue.

Category	Term	RT	Genes	Count	%	P-Value
UP_KEYWORDS	Secreted	RT		36	29.3	2.2E-10
UP_SEQ_FEATURE	signal peptide	RT		44	35.8	4.1E-8
GOTERM_CC_DIRECT	extracellular region	RT		27	22.0	6.8E-7
UP_KEYWORDS	Signal	RT		45	36.6	4.7E-6
UP_SEQ_FEATURE	disulfide bond	RT		36	29.3	6.5E-6
GOTERM_MF_DIRECT	neuropeptide hormone activity	RT		5	4.1	1.7E-5
KEGG_PATHWAY	Olfactory transduction	RT		10	8.1	1.7E-5
UP_KEYWORDS	Disulfide bond	RT		38	30.9	2.6E-5
GOTERM_BP_DIRECT	thyroid hormone generation	RT		4	3.3	3.8E-5
UP_KEYWORDS	Amidation	RT		5	4.1	1.1E-4

Figure 4.1: A screen capture of the top 10 functional annotations (ordered by descending p-value) as determined by the DAVID functional annotation tool using the important positive genes for the normal thyroid tissue class within the cancer type classification task.

Lung

The second normal class examined was lung tissue: LUSC and LUAD. Again the DAVID functional annotation tool was utilized. The genes for the LUSC and LUAD normal classes were taken from the cancer type results and combined. For these results the functional annotations were clustered and produced 71 functional clusters. The reason for using clustering in this case was because much of the top annotations in the annotation chart were related to keratinization. Keratinization is a by-product of lung distress and is common in the pathology of lung cancer patients. Since this lung tissue was all obtained from tumour-adjacent normal tissue, the presence of keratinization is expected here. Of the 71 clusters identified by DAVID, 24 of them were significantly enriched. Within these significant clusters four of them were directly related to lung function. These clusters pertained to gaseous exchange, oxygen binding, oxygen transport, and saposin proteins (involved in the pulmonary surfactant complex) [**82**]. Each of these are clear indications of the function of lung tissue and

serve to validate the results found by the model.

Breast

The third and final tissue examined was normal breast tissue. Again the DAVID annotation tool was used to produce an annotation chart. The top 14 results are presented in Figure 4.2. Fourteen results were included here because the pathways ranked between 9th and 14th are significant in normal breast tissue. The obvious pathway of significance is the one pertaining to milk proteins. A literature search revealed that keratin is also significant [**83**, **84**]. A study conducted in 1989 revealed that: "The luminal and basal epithelial cells in the human mammary gland can be distinguished in tissue sections on the basis of the pattern of keratins they express" [**83**]. Finally, the Iroquois-class homeobox proteins have been shown to be detected in breast tissue [**85**, **86**]. It should be noted that pathways for lactation and prolactin signalling were also identified by DAVID as significant, though not in the top 14 annotations. These results support that the model is able to learn genes relevant to breast tissue and its function.

Category	Term	RT	Genes	Count	%	P-Value
UP_SEQ_FEATURE	signal peptide	RT		129	37.6	1.2E-22
UP_KEYWORDS	Secreted	RT		92	26.8	5.0E-21
UP_KEYWORDS	Signal	RT		132	38.5	5.7E-16
GOTERM_CC_DIRECT	extracellular space	RT		64	18.7	1.2E-13
UP_KEYWORDS	Glycoprotein	RT		127	37.0	4.2E-11
UP_SEQ_FEATURE	glycosylation site:N-linked (GlcNAc...)	RT		121	35.3	6.5E-11
UP_SEQ_FEATURE	disulfide bond	RT		93	27.1	1.4E-10
UP_KEYWORDS	Disulfide bond	RT		103	30.0	1.7E-10
GOTERM_CC_DIRECT	extracellular region	RT		62	18.1	1.9E-9
UP_KEYWORDS	Protease inhibitor	RT		12	3.5	5.6E-6
INTERPRO	Keratin, type I	RT		7	2.0	1.3E-5
UP_KEYWORDS	Milk protein	RT		4	1.2	1.7E-5
GOTERM_CC_DIRECT	organelle membrane	RT		10	2.9	1.7E-5
INTERPRO	Iroquois-class homeodomain protein	RT		4	1.2	8.1E-5
SMART	IRO	RT		4	1.2	1.0E-4

Figure 4.2: A screen capture of the top 14 functional annotations (ordered by descending p-value) as determined by the DAVID functional annotation tool using the important positive genes for the normal thyroid tissue class within the cancer type classification task.

4.2.2 Number of Important Genes

Number of Important Genes Results

The results presented here focus on the number of positive scoring genes for each class. See Figures 4.3, 4.4, 4.5, and 4.6. Each of these figures illustrates how many genes (in blue) the model considered important for the classification of each class. The F1-scores (in red) for each class is also presented, though it is scaled by the total number of genes (26668 genes) for the sake of visualization.

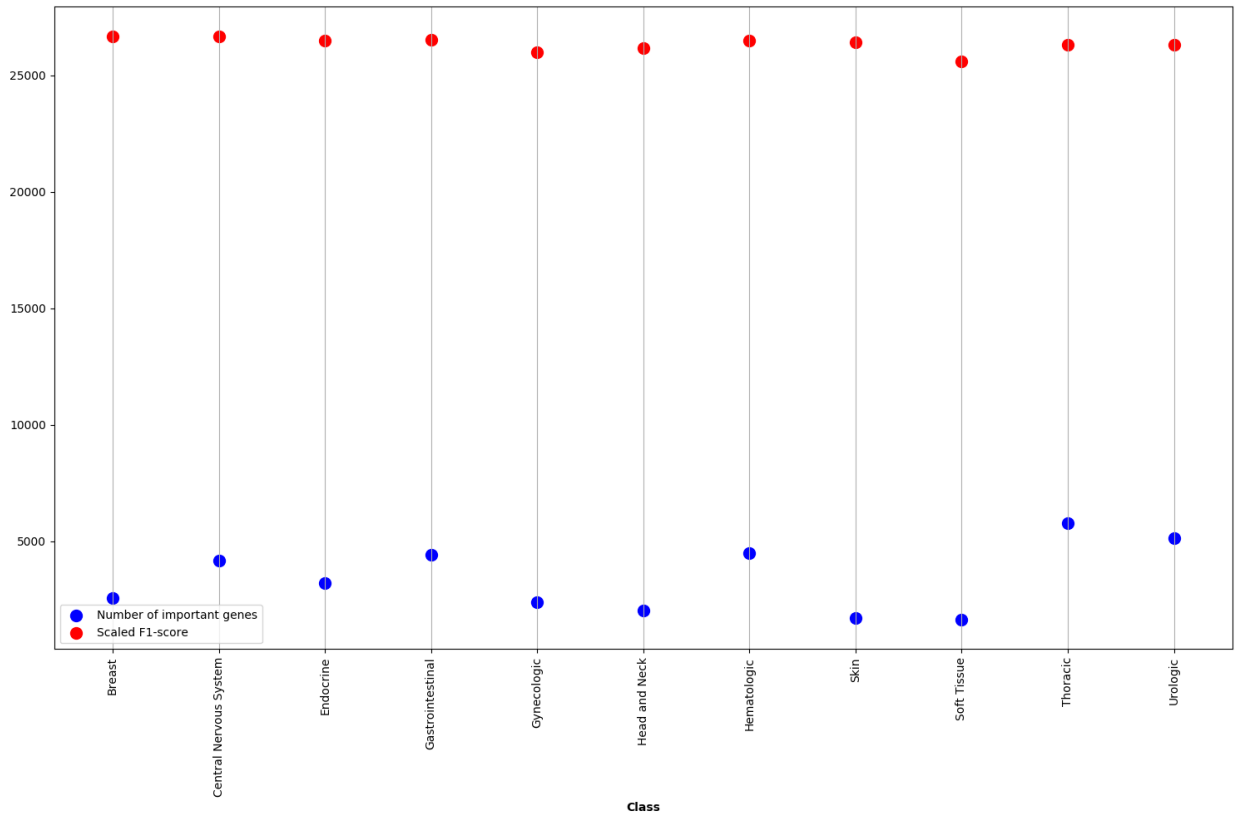

Figure 4.3: A plot showing the number of important positive genes for each class within the organ system of origin classification task in blue and the F1-score of each class in red.

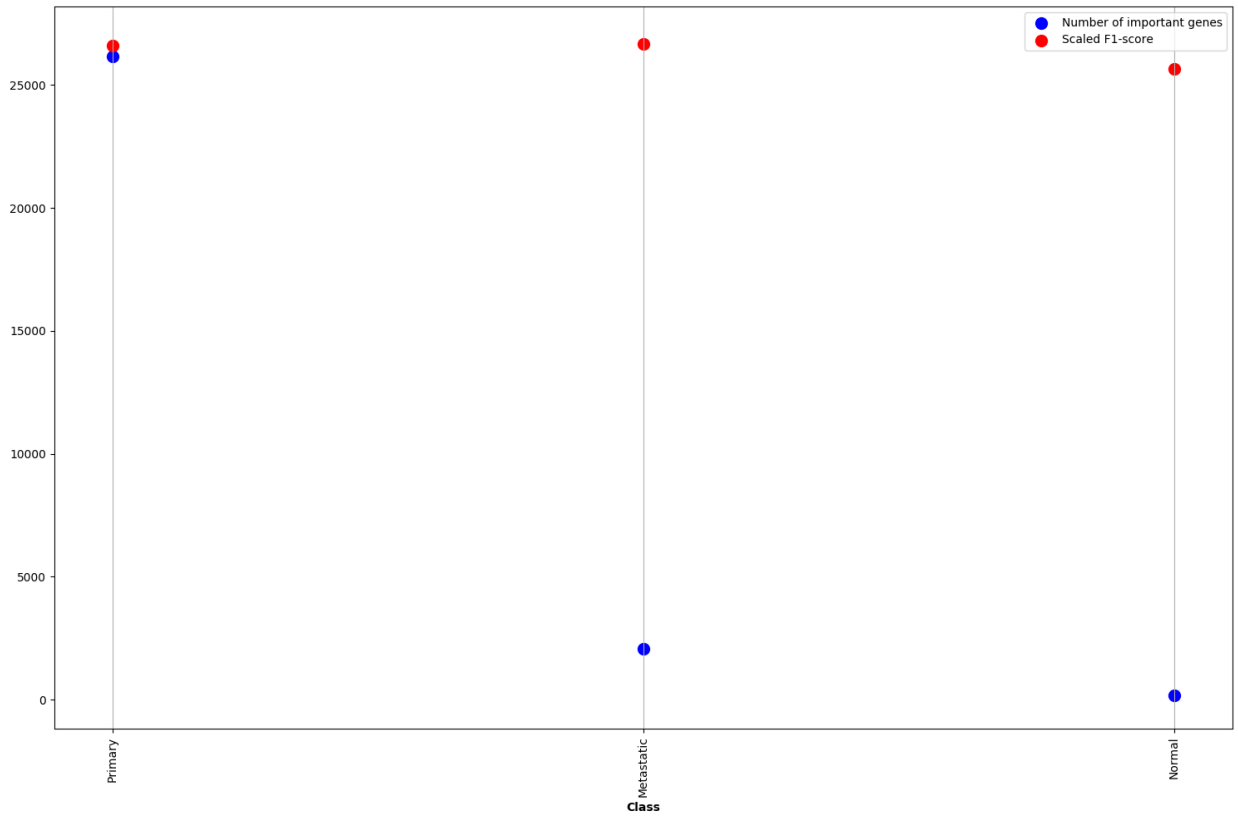

Figure 4.4: A plot showing the number of important positive genes for each class within the disease state classification task.

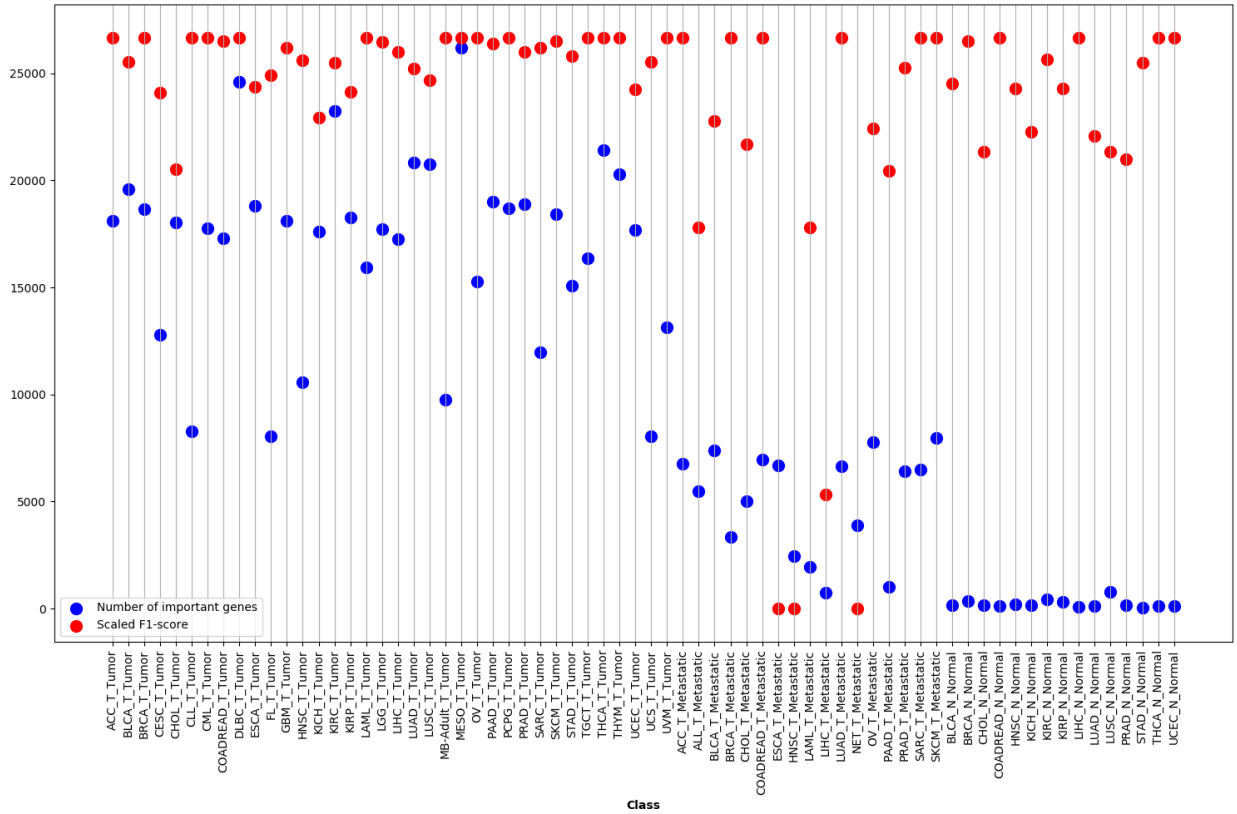

Figure 4.5: A plot showing the number of important positive genes for each class within the cancer type classification task.

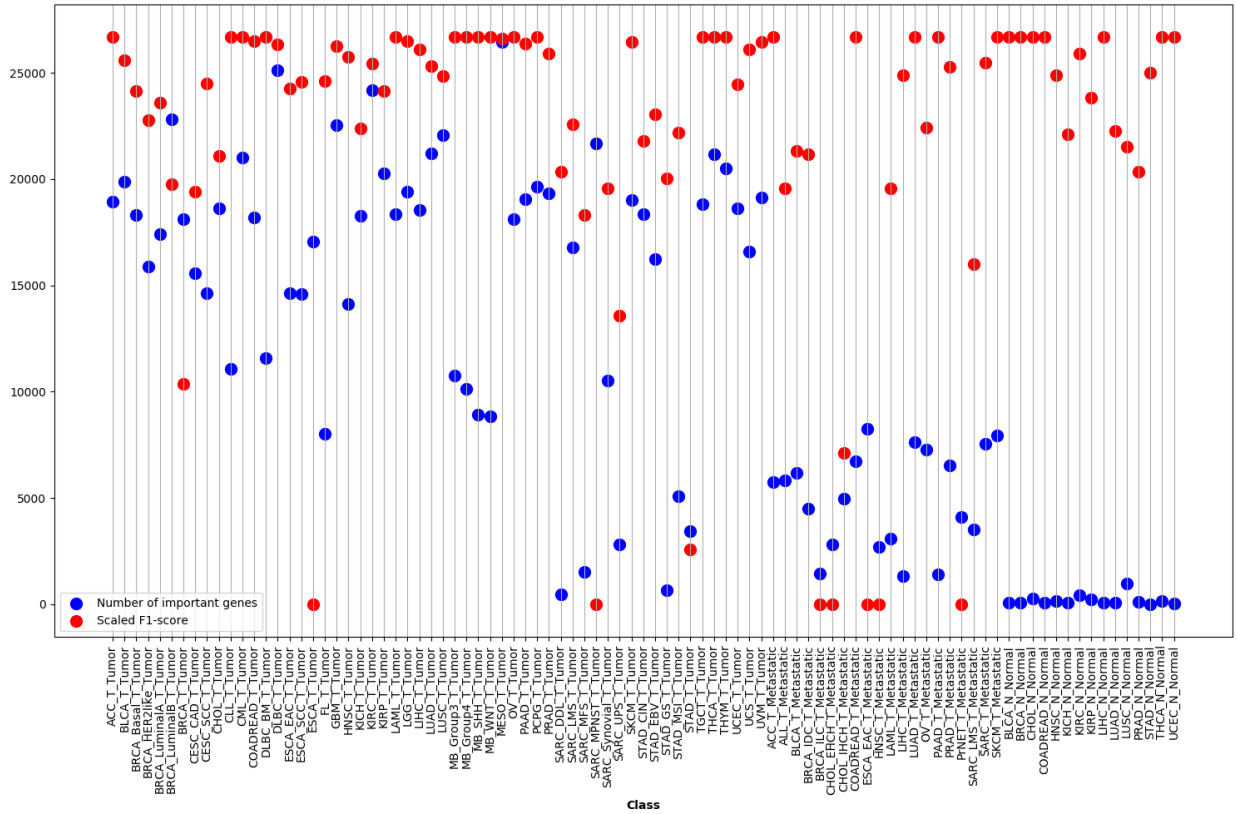

Figure 4.6: A plot showing the number of important positive genes for each class within the cancer subtype classification task.

The first notable observation from the above figures is that there appears to be three distinct plateaus in the number of important genes within the disease state, cancer type, and cancer subtype plots (Figures 4.4, 4.5, and 4.6). We see that the number of important genes is generally much higher in the primary classes (left) compared to the metastatic and normal classes (center and right respectively). The second observation is that we do not see an obvious correlation between the performance (in red) of each class and the number of important positive genes (in blue). If there was a relationship between these two values we would expect the F1-scores to also produce three tiers to reflect the tiers seen in the number of genes.

If we unpack these results further into different tasks (as seen in the above figures), we note that for the disease type classification, the primary cancer classification considers almost all of the genes as important. When compared to the approximately 2400 genes for metastatic cancers and fewer than 100 genes for normal tissues, this is a significant difference. In the cancer type task, the highest concentration of genes is around 19000 genes for primary cancers, 7500 genes for metastatic cancers and around 0 for normal tissues. A similar observation can be made at the cancer subtype level. However, the range of important genes for primary cancer subtypes increases. Notably, we see the appearance of six primary cancer subtypes that have fewer than 7000 important genes, which is much closer to the values seen within the metastatic subtypes.

The single gene with the highest importance among all classes were RPL19P12, LYVE1, PGA4, and SFTA3 in the Soft Tissue, Normal, STAD_N_Normal, and LUAD_N_Normal classes of the organ system, disease state, cancer type, and cancer subtype tasks respectively. The associated scores were 0.001563, 0.009981, 0.026148, and 0.075350. These scores indicate the percentage of the classification made using each gene for their respective classes.

Number of Important Genes Discussion

These observations beg the question of what fewer important genes mean in this context. Firstly, fewer positively important genes indicates a greater number of negatively important genes. The model has 26668 genes to consider and if it deems only 1000 as positively important for a particular class, that means there are 25668 genes that are pushing the model to not classify a sample as that class. If we consider that the performance of each class is not directly related to the number of positive important genes and the classification performance is acceptable (as we have seen in Chapter 3), then we must conclude that fewer genes indicates a sufficient compressed representation of the class for each of the specified classification tasks. In biological terms, this could mean that either fewer genes are involved, that each of the genes identified plays a more significant role than the others, or that some genes have tumour suppressing properties. As we have seen in the section on the validation of

results using normal tissues, it does appear as though the model and DeepLift have identified genes of biological relevance. This further supports the idea that the model is able to learn genes of value. Being able to classify cancers using fewer genes suggests that these cancers have a more distinct expression pattern and that some of the identified genes likely play an important role in these cancers. Through further analysis of the identified genes, we could validate existing oncogenes and potentially identify new therapeutic targets.

4.2.3 Expression Levels

This section considers if the model has identified any trends in the levels of gene expression for different cancers. Over and underexpression was determined as described in the methods section above.

Expression Levels Results

The following figures and tables show the number of over and underexpressed genes within the positive important genes for each classification task. The number of important genes, overexpressed genes, and underexpressed genes are presented in black, red, and blue respectively.

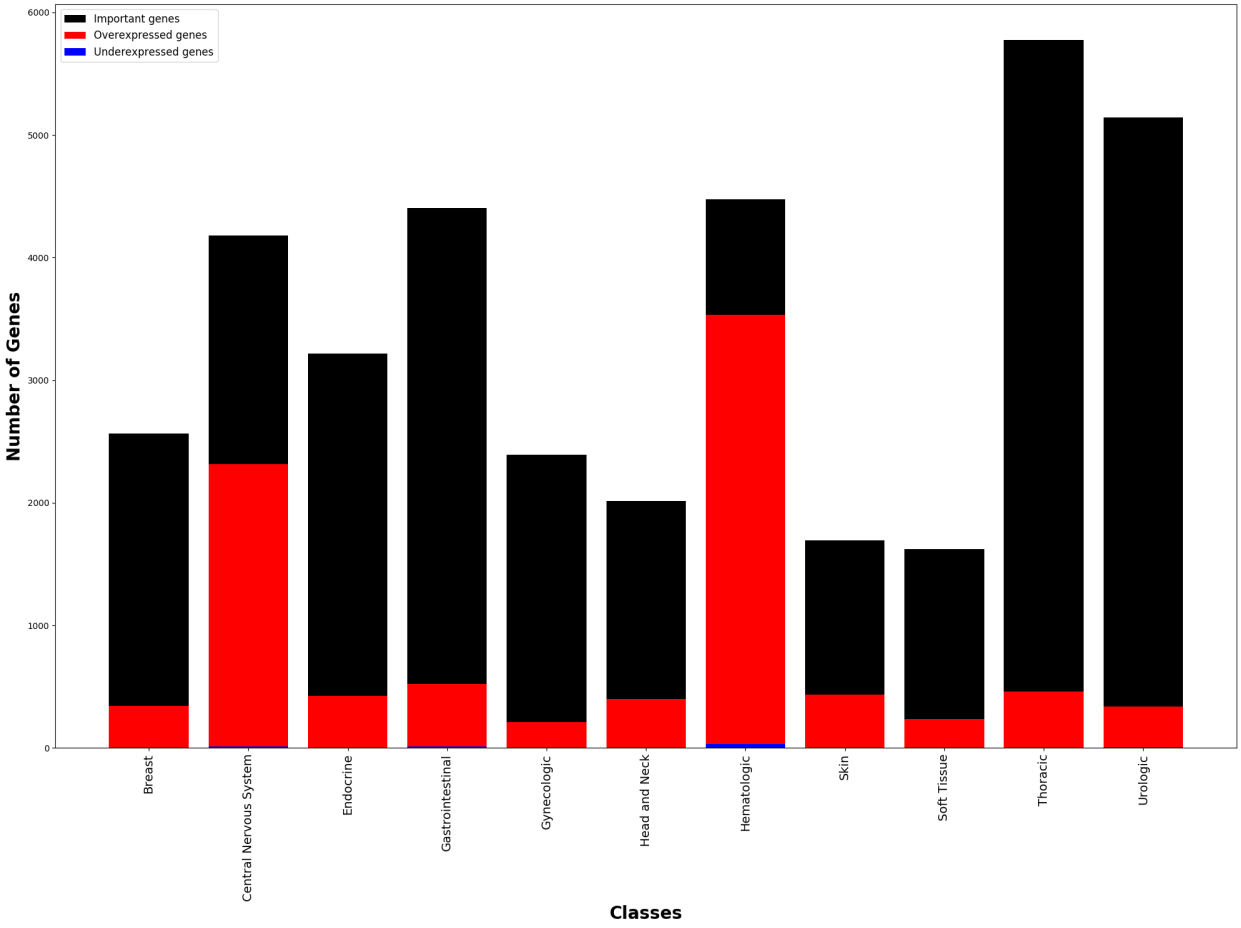

Figure 4.7: A stacked bar plot showing the number of important positive genes and the number of over and underexpressed genes for each class within the organ system of origin classification task.

Organ System of Origin	**Number of Important Positive Genes**	**Number of Overexpressed Genes**	**Number of Underexpressed Genes**
Breast	2563	341	1
Central Nervous System	4178	2306	11
Endocrine	3215	420	3
Gastrointestinal	4402	509	10
Gynecologic	2389	207	1

Organ System of Origin	**Number of Important Positive Genes**	**Number of Overexpressed Genes**	**Number of Underexpressed Genes**
Head and Neck	2012	394	3
Hematologic	4476	3500	34
Skin	1691	434	1
Soft Tissue	1620	234	0
Thoracic	5777	456	2
Urologic	5145	332	8

Table 4.2: A table listing the number of positive important genes identified by DeepLift for the organ system of origin classes along with how many of those genes are over and underexpressed.

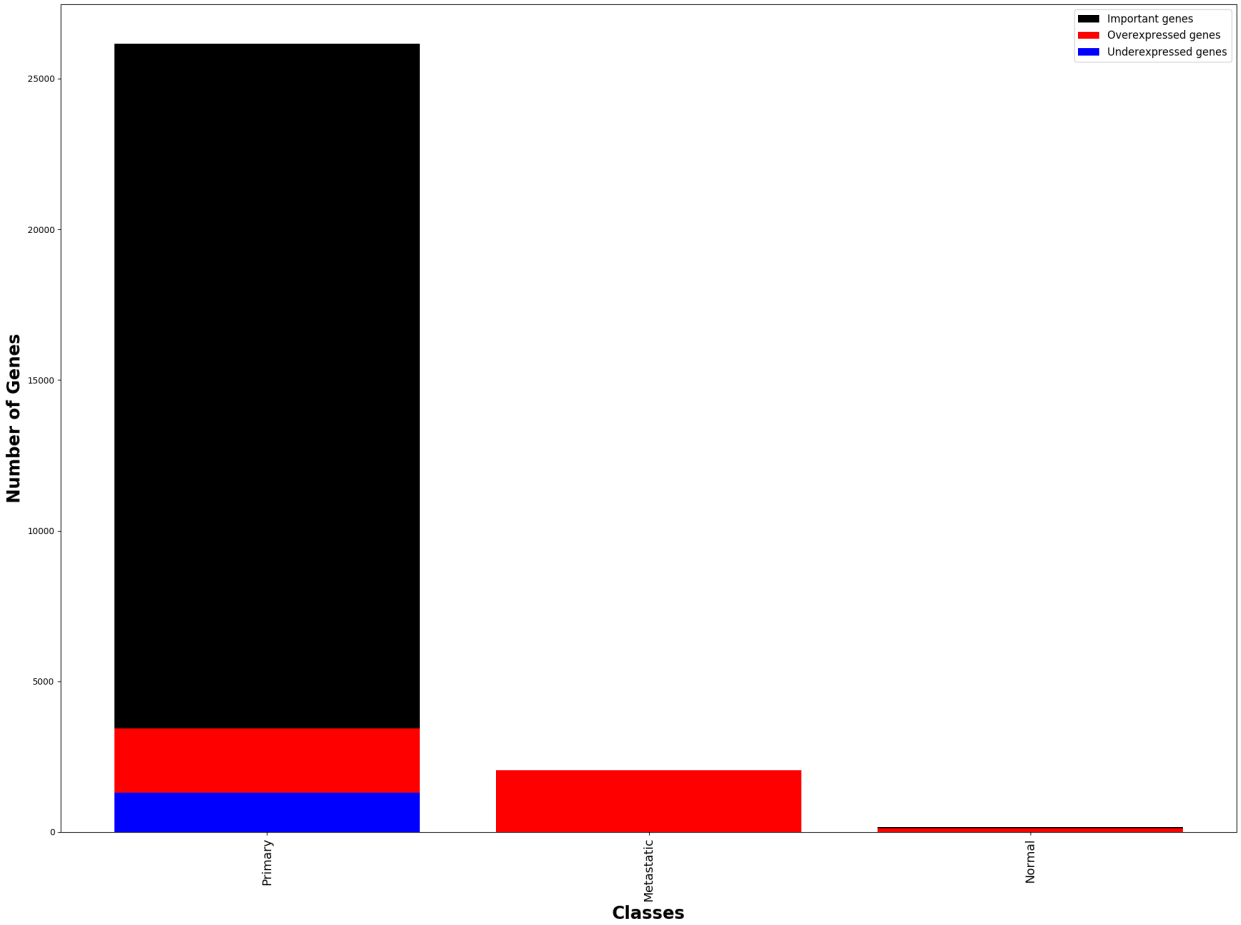

Figure 4.8: A stacked bar plot showing the number of important positive genes and the number of over and underexpressed genes for each class within the disease state classification task.

Disease State	**Number of Important Positive Genes**	**Number of Overexpressed Genes**	**Number of Underexpressed Genes**
Primary	26156	2137	1307
Metastatic	2060	2060	0
Normal	166	115	2

Table 4.4: A table listing the number of positive important genes identified by DeepLift for the disease state classes along with how many of those genes are over and underexpressed.

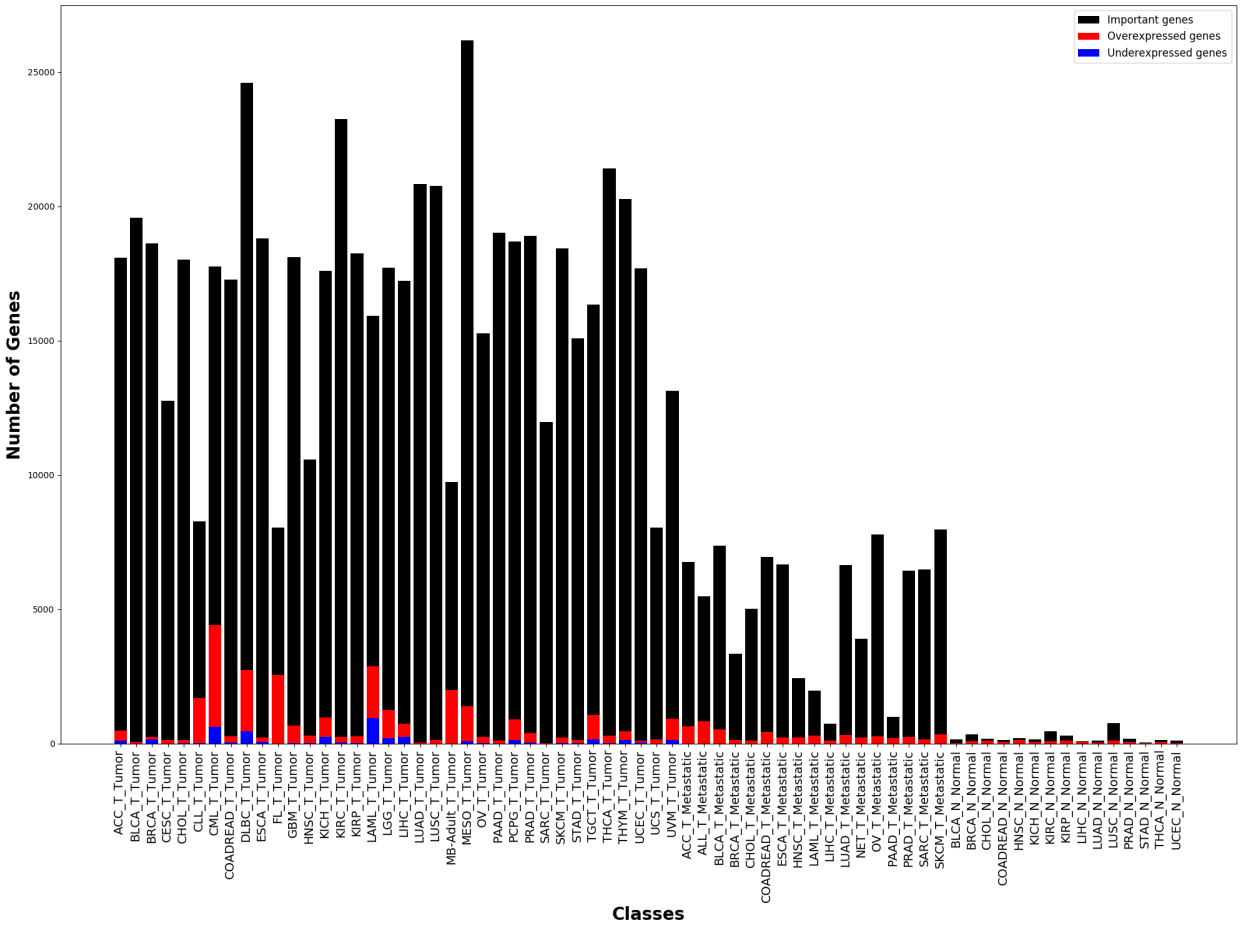

Figure 4.9: A stacked bar chart showing the number of important positive genes and the number of over and underexpressed genes for each class within the cancer type classification task.

Cancer Type	**Number of Important Positive Genes**	**Number of Overexpressed Genes**	**Number of Underexpressed Genes**
ACC_T_Tumor	18092	364	123
BLCA_T_Tumor	19575	62	9
BRCA_T_Tumor	18635	87	163
CESC_T_Tumor	12770	126	9

Cancer Type	**Number of Important Positive Genes**	**Number of Overexpressed Genes**	**Number of Underexpressed Genes**
CHOL_T_Tumor	18023	110	29
CLL_T_Tumor	8281	1684	20
CML_T_Tumor	17766	3787	628
COADREAD_T_Tumor	17277	222	53
DLBC_T_Tumor	24610	2266	466
ESCA_T_Tumor	18819	156	73
FL_T_Tumor	8048	2564	0
GBM_T_Tumor	18124	661	17
HNSC_T_Tumor	10588	298	15
KICH_T_Tumor	17595	704	263
KIRC_T_Tumor	23246	210	38
KIRP_T_Tumor	18252	241	34
LAML_T_Tumor	15919	1938	950
LGG_T_Tumor	17713	1059	205
LIHC_T_Tumor	17242	501	244
LUAD_T_Tumor	20840	46	5
LUSC_T_Tumor	20762	125	7
MB-Adult_T_Tumor	9744	2009	1
MESO_T_Tumor	26178	1311	81
OV_T_Tumor	15269	223	21
PAAD_T_Tumor	19016	119	2
PCPG_T_Tumor	18688	774	128
PRAD_T_Tumor	18903	365	39
SARC_T_Tumor	11971	22	3
SKCM_T_Tumor	18430	206	24
STAD_T_Tumor	15087	134	15

Cancer Type	Number of Important Positive Genes	Number of Overexpressed Genes	Number of Underexpressed Genes
TGCT_T_Tumor	16356	890	171
THCA_T_Tumor	21425	277	30
THYM_T_Tumor	20287	343	128
UCEC_T_Tumor	17689	81	37
UCS_T_Tumor	8040	151	6
UVM_T_Tumor	13141	790	139
ACC_T_Metastatic	6771	640	0
ALL_T_Metastatic	5476	831	0
BLCA_T_Metastatic	7372	523	0
BRCA_T_Metastatic	3340	133	0
CHOL_T_Metastatic	5022	124	0
COADREAD_T_Metastatic	6943	445	0
ESCA_T_Metastatic	6670	239	0
HNSC_T_Metastatic	2439	229	0
LAML_T_Metastatic	1970	309	0
LIHC_T_Metastatic	743	116	0
LUAD_T_Metastatic	6655	333	0
NET_T_Metastatic	3904	232	0
OV_T_Metastatic	7783	289	0
PAAD_T_Metastatic	1000	217	0
PRAD_T_Metastatic	6436	258	0
SARC_T_Metastatic	6487	159	0
SKCM_T_Metastatic	7987	342	0
BLCA_N_Normal	160	33	0
BRCA_N_Normal	354	97	0
CHOL_N_Normal	179	107	0

Cancer Type	**Number of Important Positive Genes**	**Number of Overexpressed Genes**	**Number of Underexpressed Genes**
COADREAD_N_Normal	128	79	0
HNSC_N_Normal	212	130	0
KICH_N_Normal	169	62	0
KIRC_N_Normal	454	93	0
KIRP_N_Normal	303	115	0
LIHC_N_Normal	94	78	0
LUAD_N_Normal	110	53	0
LUSC_N_Normal	771	117	0
PRAD_N_Normal	182	63	0
STAD_N_Normal	42	14	0
THCA_N_Normal	132	61	0
UCEC_N_Normal	121	29	0

Table 4.6: A table listing the number of positive important genes identified by DeepLift for the cancer type classes along with how many of those genes are over and underexpressed.

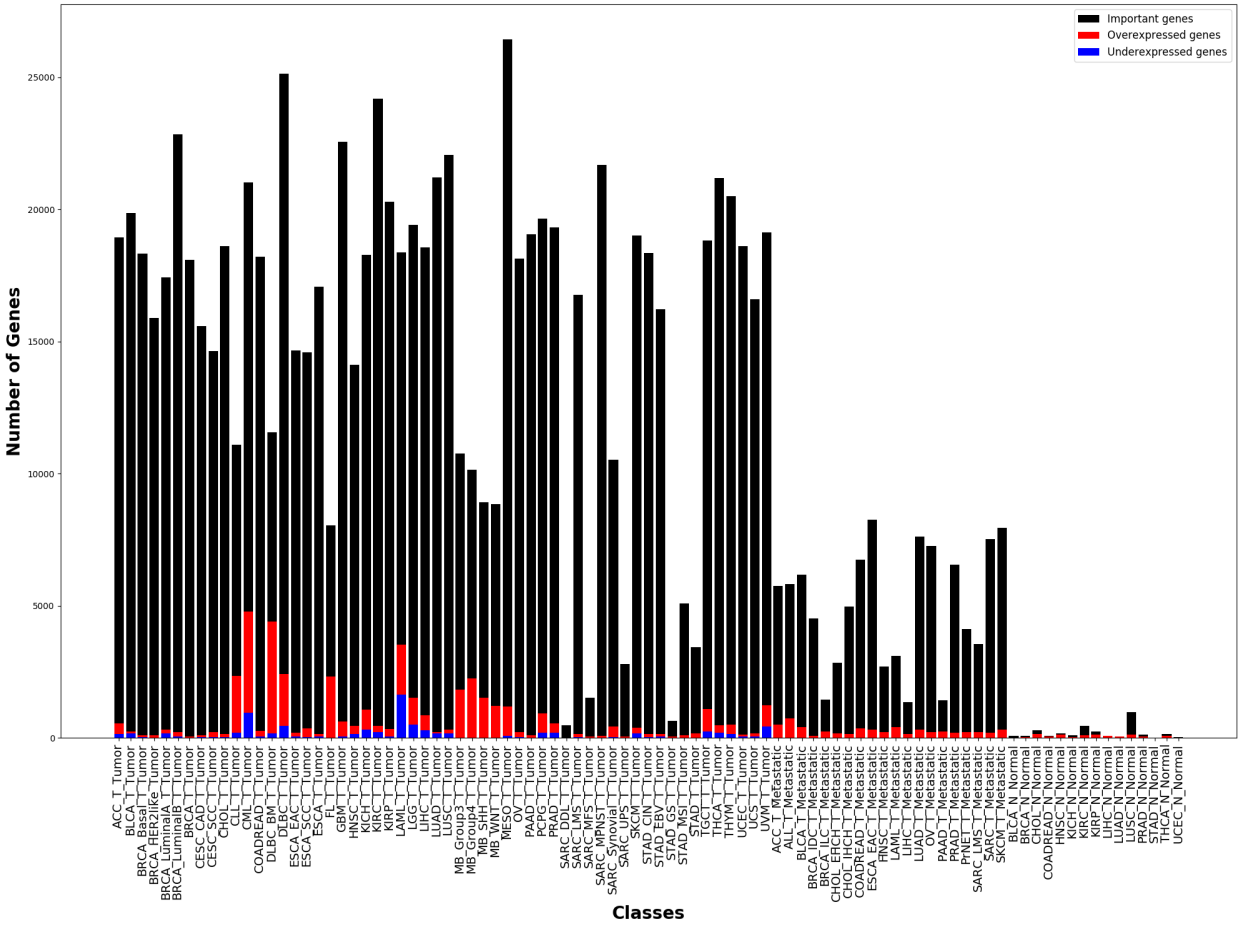

Figure 4.10: A stacked bar chart showing the number of important positive genes and the number of over and underexpressed genes for each class within the cancer subtype classification task.

Cancer Subtype	Number of Important Positive Genes	Number of Overexpressed Genes	Number of Underexpressed Genes
ACC_T_Tumor	18931	389	149
BLCA_T_Tumor	19870	69	164
BRCA_Basal_T_Tumor	18312	79	21
BRCA_HER2like_Tumor	15877	93	10
BRCA_LuminalA_T_Tumor	17415	144	164
BRCA_LuminalB_T_Tumor	22827	162	44

Cancer Subtype	Number of Important Positive Genes	Number of Overexpressed Genes	Number of Underexpressed Genes
BRCA_T_Tumor	18096	37	1
CESC_CAD_T_Tumor	15584	76	21
CESC_SCC_T_Tumor	14633	183	21
CHOL_T_Tumor	18615	118	26
CLL_T_Tumor	11088	2150	192
CML_T_Tumor	21014	3833	939
COADREAD_T_Tumor	18196	199	56
DLBC_BM_T_Tumor	11562	4224	178
DLBC_T_Tumor	25123	1954	455
ESCA_EAC_T_Tumor	14650	149	47
ESCA_SCC_T_Tumor	14587	333	21
ESCA_T_Tumor	17074	89	59
FL_T_Tumor	8038	2320	0
GBM_T_Tumor	22553	588	40
HNSC_T_Tumor	14119	305	148
KICH_T_Tumor	18276	765	300
KIRC_T_Tumor	24190	242	208
KIRP_T_Tumor	20279	296	49
LAML_T_Tumor	18362	1888	1626
LGG_T_Tumor	19417	1023	490
LIHC_T_Tumor	18558	549	296
LUAD_T_Tumor	21219	49	156
LUSC_T_Tumor	22062	135	163
MB_Group3_T_Tumor	10754	1823	11
MB_Group4_T_Tumor	10149	2249	10
MB_SHH_T_Tumor	8905	1523	0

Cancer Subtype	Number of Important Positive Genes	Number of Overexpressed Genes	Number of Underexpressed Genes
MB_WNT_T_Tumor	8838	1210	0
MESO_T_Tumor	26434	1117	62
OV_T_Tumor	18123	195	22
PAAD_T_Tumor	19057	101	0
PCPG_T_Tumor	19645	755	180
PRAD_T_Tumor	19308	358	193
SARC_DDL_T_Tumor	476	13	0
SARC_LMS_T_Tumor	16773	113	20
SARC_MFS_T_Tumor	1520	45	0
SARC_MPNST_T_Tumor	21674	59	8
SARC_Synovial_T_Tumor	10513	395	33
SARC_UPS_T_Tumor	2800	44	0
SKCM_T_Tumor	19016	204	175
STAD_CIN_T_Tumor	18355	119	22
STAD_EBV_T_Tumor	16224	108	47
STAD_GS_T_Tumor	650	53	0
STAD_MSI_T_Tumor	5089	107	0
STAD_T_Tumor	3432	167	0
TGCT_T_Tumor	18825	858	237
THCA_T_Tumor	21175	295	186
THYM_T_Tumor	20496	362	141
UCEC_T_Tumor	18618	83	47
UCS_T_Tumor	16603	119	42
UVM_T_Tumor	19136	801	423
ACC_T_Metastatic	5757	490	0
ALL_T_Metastatic	5825	729	0

Cancer Subtype	Number of Important Positive Genes	Number of Overexpressed Genes	Number of Underexpressed Genes
BLCA_T_Metastatic	6172	409	0
BRCA_IDC_T_Metastatic	4517	84	0
BRCA_ILC_T_Metastatic	1453	227	0
CHOL_EHCH_T_Metastatic	2830	156	0
CHOL_IHCH_T_Metastatic	4955	152	0
COADREAD_T_Metastatic	6747	352	0
ESCA_EAC_T_Metastatic	8244	300	0
HNSC_T_Metastatic	2704	225	0
LAML_T_Metastatic	3093	409	0
LIHC_T_Metastatic	1345	134	0
LUAD_T_Metastatic	7609	320	0
OV_T_Metastatic	7270	223	0
PAAD_T_Metastatic	1422	227	0
PRAD_T_Metastatic	6544	202	0
PrNET_T_Metastatic	4117	213	0
SARC_LMS_T_Metastatic	3537	208	0
SARC_T_Metastatic	7531	193	0
SKCM_T_Metastatic	7957	306	0
BLCA_N_Normal	62	14	0
BRCA_N_Normal	64	34	0
CHOL_N_Normal	287	150	2
COADREAD_N_Normal	65	39	0
HNSC_N_Normal	156	117	0
KICH_N_Normal	91	37	0
KIRC_N_Normal	443	87	0
KIRP_N_Normal	243	118	0

Cancer Subtype	Number of Important Positive Genes	Number of Overexpressed Genes	Number of Underexpressed Genes
LIHC_N_Normal	72	66	0
LUAD_N_Normal	59	38	0
LUSC_N_Normal	977	112	0
PRAD_N_Normal	126	56	0
STAD_N_Normal	3	0	0
THCA_N_Normal	147	67	0
UCEC_N_Normal	31	11	0

Table 4.8: A table listing the number of positive important genes identified by DeepLift for the cancer subtype classes along with how many of those genes are over and underexpressed.

Within the disease state task, the most notable observation regarding the gene expression levels of the genes identified by the model pertains to the metastatic cancer and normal tissues. If we look at Figure 4.8 and Table 4.4, we see that the model has selected 2060 genes as important for metastatic cancers and that all of them are overexpressed. Similarly, the vast majority of genes within the normal tissue class are considered overexpressed.

Within the cancer type and subtype tasks, the majority of genes across all primary classes are categorized as neither over or underexpressed. The metastatic and normal classes all utilized only overexpressed genes with the excpetion of the CHOL_N_Normal subtype in which 2 genes were underexpressed.

Expression Levels Discussion

The results above suggests that the model has found the overexpression of genes to be more informative in the context of metastatic and normal classes than of primary ones. The caveat to this is that the metastatic and normal classes both utilize many fewer genes than the primary ones and make up the minority of classes in the data set. Since the expression categories were determined using the mean expression values across all samples, the mean will be skewed towards the majority classes' gene values. So while the resultant gene lists would remain the same, redefining the boundaries of over and underexpression individually for each class would likely reveal better insight into the relevant expression levels. This becomes particularly important when trying to connect expression levels with expected gene functionality. For example, if a gene is a known tumour suppressor, carefully determining its expected expression level for a given class would allow insight into whether or not it is being underexpressed and thus driving the growth of a particular class of cancers. Given this caveat, the only real conclusion we can make of the results is that genes that are expressed above the majority of classes tend to have high importance and may contribute to needing fewer genes for classification. Until a better defined methodology for calculating the true mean expression values on a class-wise level is developed, a biological interpretation of these results should be reserved.

4.2.4 Enriched Pathways: Metastatic Cancer Disease State

DAVID Functional Annotation Chart Results

Figure 4.4 and Table 4.4 show that the disease state task identified 2060 genes contribute positively to the classification of metastatic cancers by the model. These genes were input into the DAVID functional annotation tool to identify enriched functional pathways [**76**, **77**].

The DAVID functional annotation chart returned 163 records and the top 10 results are presented in Figure 4.11. This figure is ordered by descending p-value.

Category	Term	RT	Genes	Count	%	P-Value
INTERPRO	Immunoglobulin V-set	RT		58	3.0	3.1E-25
SMART	IGv	RT		43	2.2	8.6E-24
INTERPRO	Immunoglobulin-like domain	RT		69	3.5	4.3E-19
INTERPRO	Immunoglobulin-like fold	RT		69	3.5	2.5E-14
KEGG_PATHWAY	MicroRNAs in cancer	RT		35	1.8	3.5E-14
UP_KEYWORDS	Ribosomal protein	RT		23	1.2	5.9E-9
GOTERM_MF_DIRECT	structural constituent of ribosome	RT		22	1.1	2.3E-8
KEGG_PATHWAY	Ribosome	RT		18	0.9	6.7E-8
GOTERM_CC_DIRECT	ribosome	RT		19	1.0	7.5E-8
GOTERM_BP_DIRECT	translation	RT		22	1.1	2.3E-7

Figure 4.11: A screen capture of the top 10 functional annotations (ordered by descending p-value) as determined by the DAVID functional annotation tool using the important positive genes for the metastatic class within the disease state classification task.

DAVID Functional Annotation Chart Discussion

The first observation to note is that within the top annotations we see 'MicroRNAs in cancer'. According to KEGG, the pathway highlighted here corresponds to 'a cluster of small non-encoding RNA molecules of 21 - 23 nucleotides in length, which controls gene expression post-transcriptionally either via the degradation of target mRNAs or the inhibition of protein translation' [**79**, **80**, **81**]. This corresponds with the information presented in Chapter 1 about the role of microRNA in tumourigensis. Furthermore, studies have suggested that specific miRNAs play a role in metastatic cancers [**87**, **88**]. The fact that this is a highly enriched pathway for this class is a promising result that is supported by the literature. Through further in depth analysis, we may be able to identify metastatic-specific miRNA genes that could make suitable therapeutic targets.

The second observation to note is that four of the top 10 functional annotations are related to the ribosome in some way. Studies have indicated that changes in the regulation of ribosomal proteins can be associated with poor prognosis for cancer patients [**89**, **90**, **91**, **92**]. Additionally, increased expression of ribosomal RNAs have been correlated with the development of some cancer types and in some cases poor prognosis and/or metastasis. [**89**, **93**, **94**, **95**, **96**]. The prevalence of genes related to ribosomal function further supports the evidence presented in the literature and further validates the biological significance of the identified genes. Ribosomal function plays a key role in the development of cancers and the through careful examination of the genes identified here, ribosomal genes could be further explored as therapeutic targets [**97**, **98**].

Finally, the top four annotations strongly suggests that immunoglobulin (IG) is important to classifying metastatic cancers. In particular, the V-set of immunoglobulin is listed twice with the highest p-values. This set of immunoglobulin has been found to be overexpressed in and indicative of poor prognosis for patients with advanced gastric cancers [**99**]. If evidence suggests increased expression of V-set IG is found to prognosticate advanced cancers, then IG should be explored for causal links to metastatic cancers, as they are themselves are an advanced form of the disease. Furthermore, a second study also validates the importance of the IG family of genes (which includes the V-set) and found that dysregulated expression of IG shows prognostic value for breast cancers [**100**]. We can speculate that the model has learned to detect changes in the expression of IG genes and utilizes it to inform the classification of metastatic cancers.

DAVID Functional Annotation Clustering Results

The DAVID functional annotation clustering tool was used to cluster the functional annotations for the metastatic cancer genes within the disease state task and returned 53 clusters. Of the 53 clusters only nine had a significant enrichment score (above 1.3) [**76**]. The top three enriched pathways include groups involving immunoglobulin, ribosomal translation,

and mitchondrial translation/mitochondrial ribosome pathways with enrichment scores of 12.58, 5.79, and 2.71 respectively.

DAVID Functional Annotation Clustering Discussion

The clustering results further support the important role of IG and ribosomes discussed in the preceding subsection. The role of mitochondria was not discussed earlier in this thesis but has been shown to play a role in the formation of cancers and metastases [**101**]. In particular, a number of mitochondrial ribosomal proteins (mitoribosomal proteins or MRPs) have been implicated in the development of various metastatic cancers [**101**, **102**]. The presence of all three of these annotation groups, both in the cancer literature and in the results presented here, are excellent further indicators of the ability for the model to identify important biological features of metastatic cancer. The implication of each of these pathways should be investigated further.

Summary of Enriched Pathways for Metastatic Cancer in the Disease State

The results presented above have indicated that immunoglobulin, microRNA, and the ribosome play significant roles in the classification of metastatic cancers within the disease state task. Furthermore, the scientific literature seems to indicate that there is some validity to this observation made by the model. The value of this insight is that it was made within the disease state task and thus applies across multiple metastatic cancer types. The list of pathways identified within this task could be further examined and exploited to try and better understand common characteristics of metastatic cancers as a whole. One caveat to consider with these results is the implication of batch effect as a result of all the metastatic cancers being from a single data source external to the bulk of the training data (see the section on batch effect below).

4.2.5 Enriched Pathways: Primary Cancer in the Disease State Task

The genes identified as important for the primary cancer class contains almost all of the total available genes (26668 genes). In this case, it would be uninformative to use the full gene list for enriched pathway analysis. However, since each gene is given a score, it is possible to rank the importance of each gene as a percentage of the total classification decision. Therefore, enriched pathway analysis results are presented using the top 25% of genes ranked by descending importance score. Note that individual cancer type and subtype results will not be presented but rather the focus will remain on the more general categories of primary and metastatic cancers.

DAVID Functional Annotation Chart Results

The top 25% of important genes for primary cancer within the disease state task consists of 2780 genes. These genes were input into DAVID and the functional analysis chart tool was utilized to generate the results presented in Figure 4.12.

Category	Term	RT	Genes	Count	%	P-Value
UP_SEQ_FEATURE	chain:Cancer/testis antigen 47A	RT		11	0.5	5.2E-11
UP_KEYWORDS	DNA repair	RT		45	2.2	4.0E-6
UP_KEYWORDS	DNA damage	RT		51	2.5	4.6E-6
GOTERM_BP_DIRECT	DNA repair	RT		39	1.9	6.3E-6
UP_KEYWORDS	Alternative splicing	RT		851	41.7	2.3E-5
UP_SEQ_FEATURE	domain:Ig-like V-type	RT		25	1.2	2.6E-5
UP_KEYWORDS	rRNA processing	RT		19	0.9	4.1E-5
KEGG_PATHWAY	RNA transport	RT		29	1.4	5.1E-5
GOTERM_CC_DIRECT	spliceosomal complex	RT		20	1.0	5.1E-5
GOTERM_CC_DIRECT	nucleoplasm	RT		259	12.7	5.2E-5

Figure 4.12: A screen capture of the top 10 functional annotations (ordered by descending p-value) as determined by the DAVID functional annotation tool using the top 25% of important positive genes for the primary class within the disease state classification task.

DAVID Functional Annotation Chart Discussion

Looking at the results in Figure 4.12, we see that the most significant pathway listed is cancer-related. The cancer/testis antigens are a group of proteins that are normally expressed only in testicular germ cells but are found to be expressed in numerous cancers [**103**]. These antigens include a number of other types of genes including GAGE, MAGE, and BAGE [**103**]. GAGE genes were found to also be significant within the primary cancer results and are discussed below in the context of functional annotation clustering. Another notable result in the top 10 enriched pathways presented in Figure 4.12, is the V-type IG-like pathway. We have seen that the V-set IG pathway was important for the metastatic cancer class (see above) and the results presented here, along with the literature presented above, support the significance of this pathway in cancers in general.

DAVID Functional Annotation Cluster Results

The DAVID functional annotation clustering tool returned 13 significant clusters out of 198 identified functional clusters. The top three clusters had enrichment scores of 5.31, 3.41, and 2.94 corresponding to functional pathways involving DNA repair/damage, G antigens (GAGE), and putative proteins (see Figure 4.13).

Annotation Cluster 1	Enrichment Score: 5.31			Count	P_Value
UP_KEYWORDS	DNA repair	RT		45	4.0E-6
UP_KEYWORDS	DNA damage	RT		51	4.6E-6
GOTERM_BP_DIRECT	DNA repair	RT		39	6.3E-6
Annotation Cluster 2	**Enrichment Score: 3.41**			**Count**	**P_Value**
UP_SEQ_FEATURE	chain:G antigen 5	RT		8	2.8E-4
UP_SEQ_FEATURE	chain:G antigen 6	RT		8	2.8E-4
UP_SEQ_FEATURE	chain:G antigen 7	RT		8	2.8E-4
UP_SEQ_FEATURE	chain:G antigen 8	RT		8	2.8E-4
UP_SEQ_FEATURE	chain:G antigen 1	RT		8	2.8E-4
UP_SEQ_FEATURE	chain:G antigen 12C/D/E	RT		8	2.8E-4
UP_SEQ_FEATURE	chain:G antigen 12F/G/I	RT		8	2.8E-4
UP_SEQ_FEATURE	chain:G antigen 2A/2B	RT		8	2.8E-4
UP_SEQ_FEATURE	chain:G antigen 3	RT		8	2.8E-4
UP_SEQ_FEATURE	chain:G antigen 4	RT		8	2.8E-4
SMART	SM01379	RT		10	5.5E-4
INTERPRO	GAGE	RT		8	6.3E-3
Annotation Cluster 3	**Enrichment Score: 2.94**			**Count**	**P_Value**
UP_SEQ_FEATURE	chain:Putative protein FAM90A10	RT		6	1.2E-3
UP_SEQ_FEATURE	chain:Putative protein FAM90A18/FAM90A19	RT		6	1.2E-3
UP_SEQ_FEATURE	chain:Putative protein FAM90A20	RT		6	1.2E-3
UP_SEQ_FEATURE	chain:Putative protein FAM90A24	RT		6	1.2E-3
UP_SEQ_FEATURE	chain:Putative protein FAM90A7	RT		6	1.2E-3
UP_SEQ_FEATURE	chain:Putative protein FAM90A8	RT		6	1.2E-3
UP_SEQ_FEATURE	chain:Putative protein FAM90A9	RT		6	1.2E-3

Figure 4.13: A screen capture of the top 3 functional annotation clusters (ordered by descending enrichment score) as determined by the DAVID functional annotation cluster tool using the top 25% of important positive genes for the primary class within the disease state classification task.

DAVID Functional Annotation Cluster Discussion

The first functional cluster identified by the clustering tool pertains to DNA repair and damage. This result is somewhat difficult to reconcile. Numerous studies report that the underexpression of DNA repair genes is associated with an increased likelihood of tumourgenesis as a result of increased genomic instability [**104**, **105**, **106**]. However, underexpression of DNA repair genes in patients already afflicted with cancer is associated with poorer prognoses and treatment outcomes [**106**]. The consensus seems to contradict the results shown here. We would expect to see overexpression of DNA repair genes in the context of primary cancers and underexpression, if at all, in the metastatic cancers.

The second functional cluster pertains to G antigens (GAGE). GAGE genes have been found to be upregulated across numerous cancers and support the importance placed on this functional cluster by the model [**107**, **108**, **109**]. They are expressed in response to epigenetic dysregulation in cancer cells but are otherwise inactive [**107**, **108**, **109**, **110**]. The only exceptions to this are during the developmental period and within testicular germ cells [**107**, **108**, **109**, **110**]. The exact mechanism by which GAGE genes impact tumourigenesis is unclear but they are being are explored as potential therapeutic targets [**107**, **108**, **109**, **110**]. It should be noted that GAGE genes are within the same category of genes as the cancer/testis antigens discussed above, further encouraging their relevance within the model.

The third functional cluster involves a set of putative proteins. By nature of being putative, we cannot speculate on the value or function of these genes. However, these genes could be noted for future experimentation to determine their functionality where possible.

Summary of Enriched Pathways for Primary Cancer in the Disease State Task

Generally, the list of functional annotations for primary cancers was less informative of the underlying biology than within the metastatic cancer class. This outcome was expected given that the model utilized almost all of the genes to make a primary cancer classification within the disease state task. The high gene usage results in each gene's contribution being significantly reduced and thus having weaker importance. This increased gene contribution confounds the resulting functional annotations as 75% of the genes identified were excluded from the results presented here. Using only 25% of genes is effectively an arbitrary cut-off point and may not have any real underlying biological significance. Given the vast number of genes used, these are simply genes ranked slightly above the others.

4.2.6 RNA Genes

The results presented in this section serve to quantify and present the role that RNA genes play in different cancers. The majority of the discussion that follows will focus on the disease state and cancer type classification tasks as the trends remain similar within the cancer subtype task. These tasks sufficiently exemplify the larger trends across cancers. It should be noted that there are 2890 RNA genes in the full set of genes available to the model and that this comprises 10.8% of the available genes.

Organ System of Origin Task Results

Figure 4.14 presents the proportion of RNA genes identified as important within each class of the disease state task.

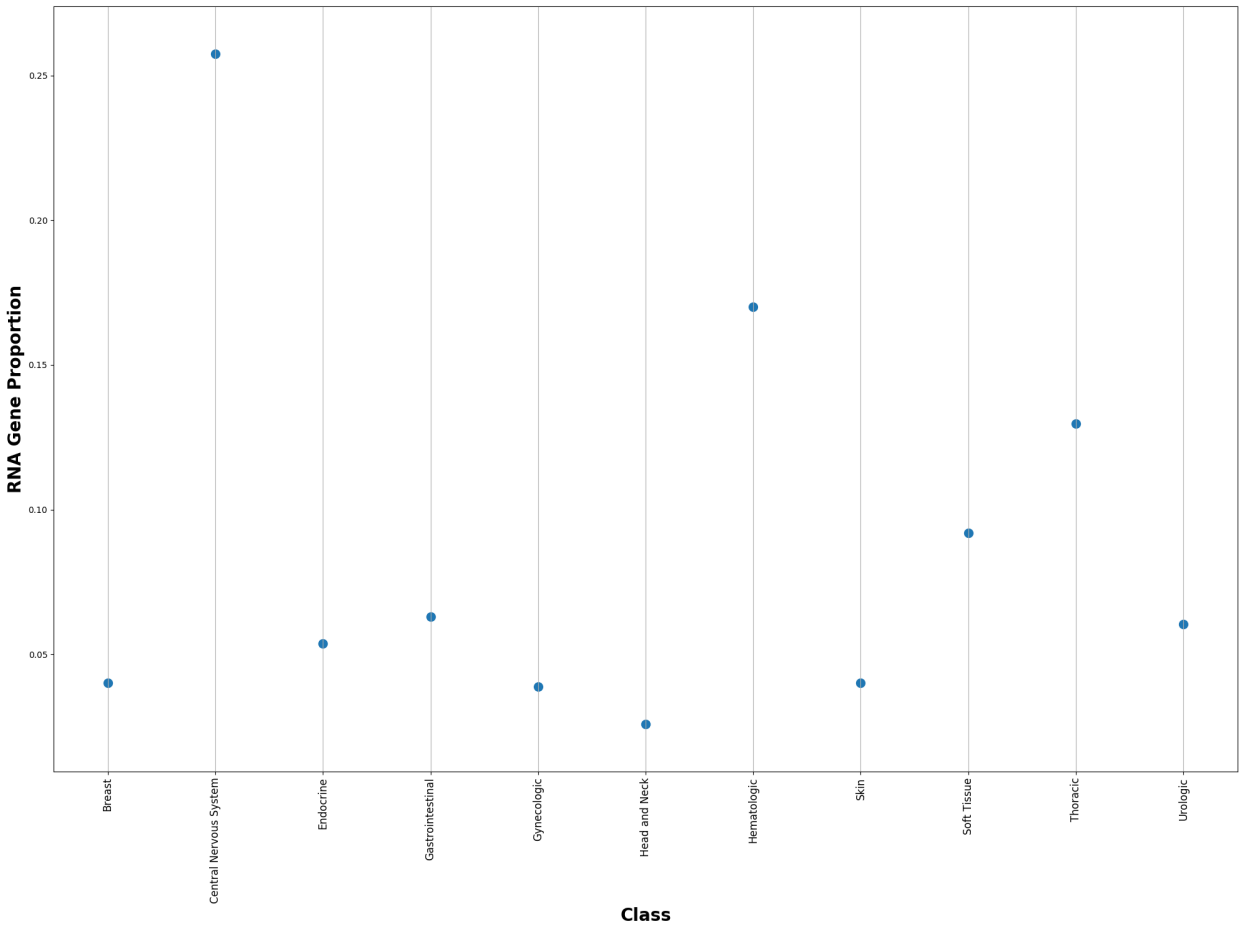

Figure 4.14: A scatter plot showing the proportion of RNA genes within the positive important genes identified for the organ system of origin classes.

We note that the highest RNA gene proportion is found in the central nervous system, hematologic, thoracic and soft tissues classes with values of 0.26, 0.17, 0.13, and 0.095 respectively. The other classes have values ranging from 0.025 to 0.053.

Organ System of Origin Task Discussion

The two highest RNA gene proportions are found in the central nervous system and hematologic classes. When we consider the RNA gene proportions of cancer types with the highest proportions (see sections below) we find that the top three are within cancer types

of the central nervous system and hematologic organ systems (CLL, FL, and MB). Since these organ systems already include higher RNA gene proportions in their important genes, it may have contributed to the high RNA gene proportion found in the related cancer types.

Disease State Task Results

Figure 4.15 presents the proportion of RNA genes identified as important within each class of the disease state task.

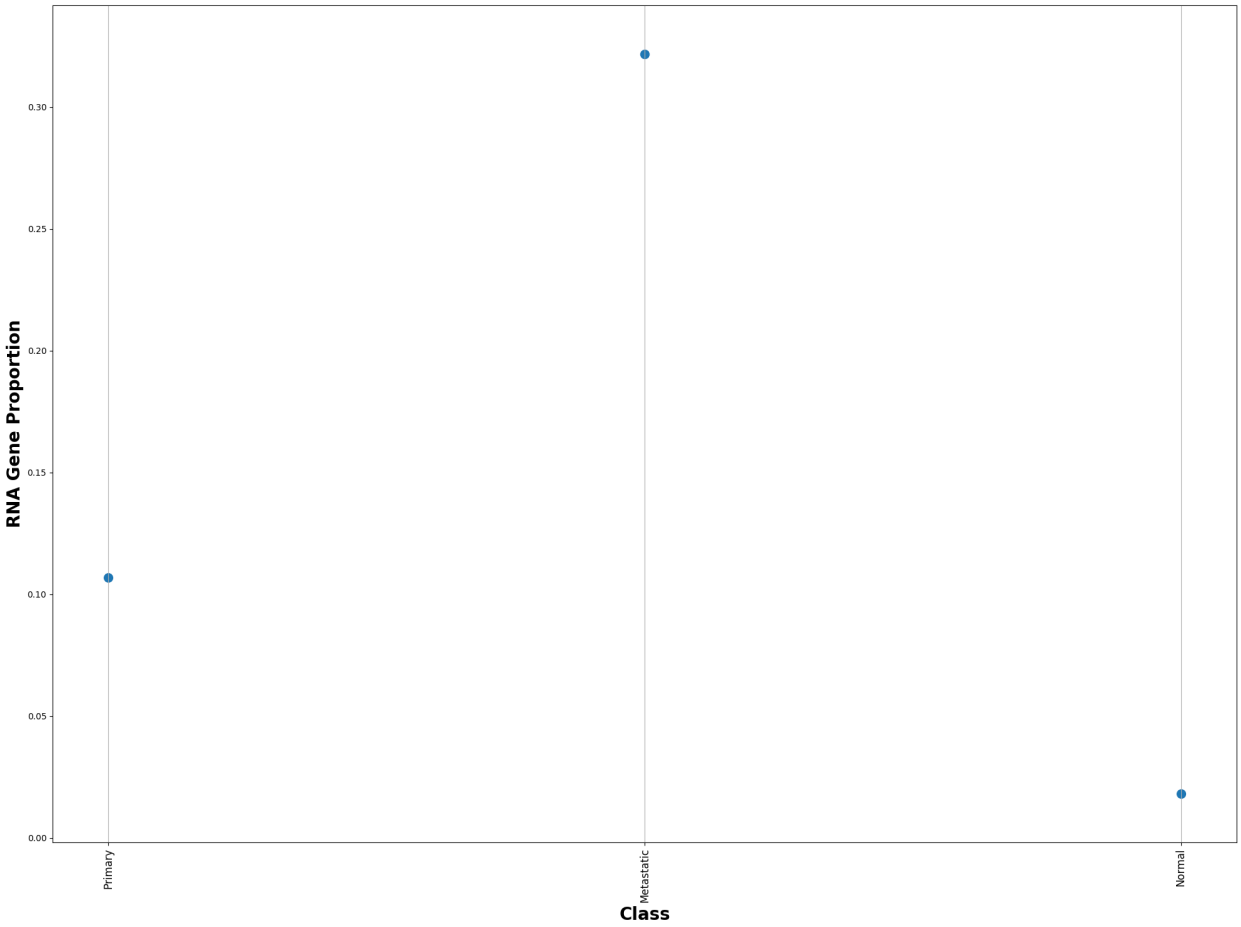

Figure 4.15: A scatter plot showing the proportion of RNA genes within the positive important genes identified for the disease state classes.

Figure 4.15 illuminates a significant difference between the number of RNA genes deemed important for metastatic classes when compared to primary and normal ones. We note that the RNA gene involvement is approximately three times as high in the metastatic class as in the primary one with proportions of 0.32 and 0.11 respectively. We also note that RNA gene importance is approximately 0.02 within the normal class.

Disease State Task Discussion

The small number of RNA genes utilized in the normal class coincides with the small number of important genes seen in Section 4.2.2. This suggests that the model has learned to ignore the majority of genes, RNA or otherwise, for normal tissues. In fact, the model has deemed most genes as an indication of non-normal classes. This implies that there are very few genes that are not important to the classification of cancer within the context of this model.

With regards to RNA genes in the primary cancer class, it is important to remember that the model has deemed almost all of the genes available (see Figure 4.4) as important. As a result, the number of RNA genes closely reflects the number available within the entire gene set (10.8% RNA genes). While this fact reduces the value of examining the RNA gene proportion within the primary class, the opposite is true of the metastatic class.

Given the otherwise high gene exclusion rate within the metastatic class, the fact that the model has elected to deem such a high proportion of RNA genes as important is significant. This suggests that RNA genes have a very strong impact on the classification of metastatic samples within the context of the disease state learning task and accounts for almost one third of a classification decision. When conducting further analysis on the genes highlighted for metastatic cancers special attention should be made to consider the interaction between RNA and non-RNA (such as miRNA) genes. It may be possible to look for correlations between the expression patterns of RNA genes and their related coding genes.

It is worth noting here that the disease state task has the highest potential for being negatively affected by batch effect. Since all of the metastatic cancers are from a single and different data source than the bulk of the training data (TCGA), batch effect can pose a serious issue for biological interpretation. At this level of classification task, the susceptibility to learning how to simply differentiate data sources is high and thus any biological interpretation of results should consider this implication. Batch effect is discussed in more detail in Section 4.3.

Cancer Type Task Results

Figure 4.16 presents the proportion of RNA genes identified as important within each class of the cancer type task.

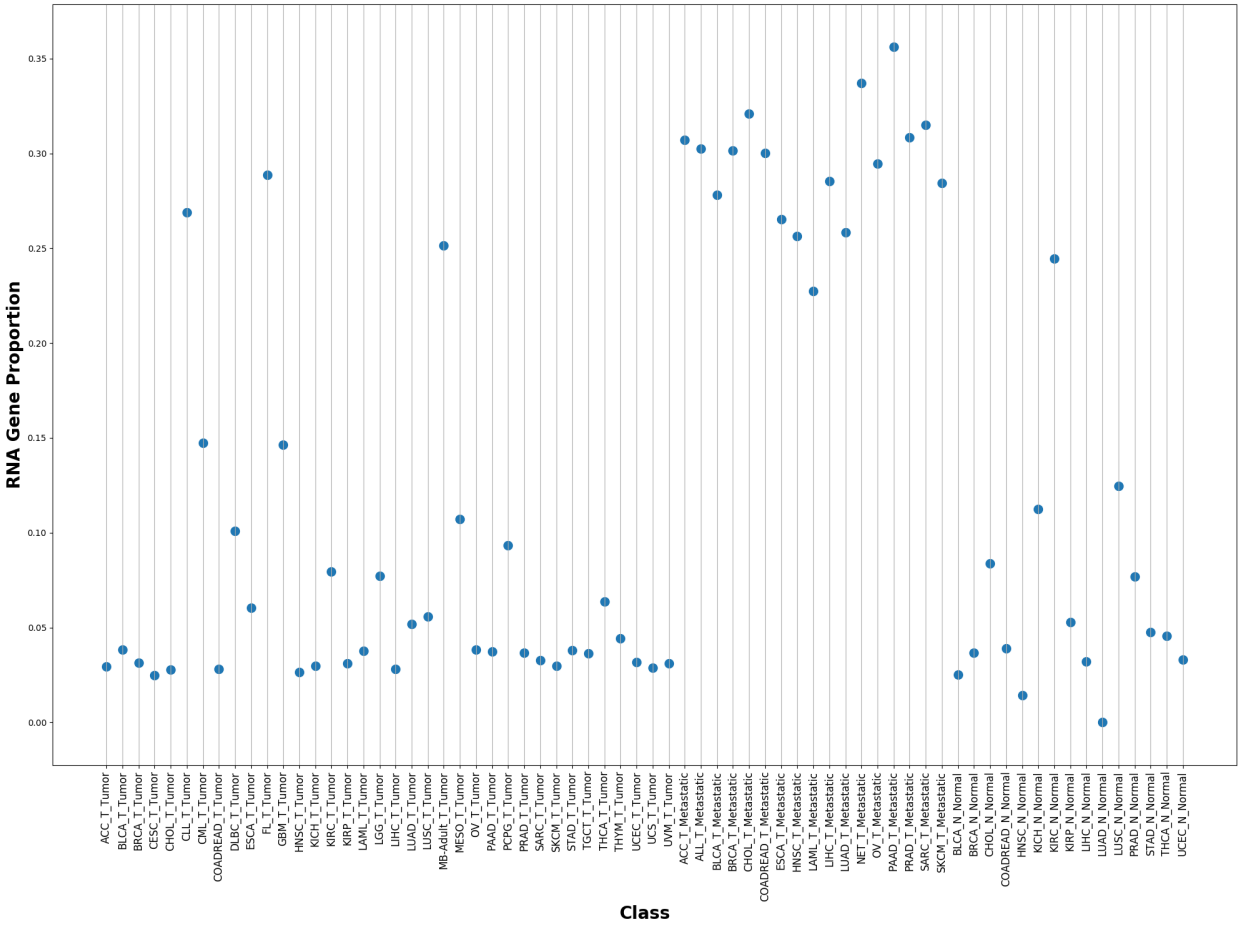

Figure 4.16: A scatter plot showing the proportion of RNA genes within the positive important genes identified for the cancer type classes.

In Figure 4.16, we see that the proportion of RNA genes utilized in the metastatic cancer types range from 0.23 to 0.37 and are spread relatively evenly throughout this range. The range for primary cancer types is 0.025 to 0.275 with the vast majority having an RNA proportion of approximately 0.03. When comparing these two proportions, the vast majority of primary cancers use 7 times fewer RNA genes than metastatic cancers. The normal tissue classes share a similar range to that of the primary cancers with proportions from near 0.0 and 0.245. The KIRC normal presents as the largest outlier with an RNA gene proportion of 0.25.

The metastatic cancer types with the lowest and highest proportion of RNA genes are LAML and OV respectively. As a whole, the metastatic cancer types proportions compose one cluster. Within the primary cancers, there are three classes whose RNA gene proportions are within the range of metastatic cancers, making them outliers within the primary class. These classes are CLL, FL, and MB, and they have RNA proportions of 0.265, 0.29, and 0.25 respectively.

Cancer Type Task Discussion

Given that the total gene set available to the model contains 10.8% RNA genes, the fact that the majority of normal and primary cancer type classes rely on fewer than 4% RNA genes suggests that RNA genes were selected against. Within metastatic cancer classifications we saw the opposite effect with high proportions of RNA gene involvement. We therefore have not only evidence of RNA gene importance in metastatic classes, but evidence of RNA gene aversion in primary and normal ones. Combined, these observations strongly suggest that RNA genes play a much more significant role in the classification of metastatic cancers when compared to primary cancers and normal tissues.

Regarding the normal tissues, there is no easily observable correlation between the number of RNA genes utilized for the corresponding primary cancer type. For example, we see that the two lung normal classes, LUAD and LUSC, have both the highest and lowest RNA gene usage within the normal classes. Their corresponding primary cancers both show only slightly elevated RNA gene usage with proportions just above 0.05. This indicates that the model is able to detect and learn the differences in RNA expression patterns between normal and primary cancer tissues. We can go one step further and speculate that, since the normal classes are defined by their adjacent cancer types, perhaps the RNA expression patterns differ between lung tissues that have developed adenocarcinoma verses squamous cell carcinoma. One study by Shi et al. (2014) has found 2961 microRNAs that are differentially expressed between lung cancers and normal lung tissue [**111**]. Following up on this work, another

study by Venugopal et al. (2019) detected fundamental differences in the gene expression patterns between lung adenocarcinoma and squamous cell carcinoma [**112**]. It is plausible that given the high number of microRNA genes implicated in lung cancer and the differential gene expression between lung cancer types, that at least some of these changes in expression are driven by non-coding genes. Further functional analysis of the RNA genes identified in each class may serve to validate this speculation.

There are two aspects of these results that should be further analyzed. The first aspect is what the increased RNA presence in metastatic cancers means in terms of biology, and the second involves examining the RNA proportions for the outlying primary cancers. These will be discussed in the subsections below.

Cancer Type Task Discussion: RNA Genes in Metastatic Cancers

The high level of RNA gene importance in metastatic cancer types indicates that these genes play a significant role in differentiating metastatic cancers from primary ones. The question is whether or not there is biological significance to this. Scientific evidence is beginning to suggest that non-coding RNA plays an important role in regulating the developmental transitions of cells [**113**]. In particular, the epithelial to mesenchymal transition (EMT) is a key developmental transition that is indicated at the start of metastasis. EMT is the mechanism by which cells can reactivate embryonic morphogenesis and ultimately contributes to the ability of cells to propagate and migrate to distant organ systems [**113**]. Non-coding RNA, as they pertain to cancer, are also implicated in the disruption of the cellular signalling pathways involved in the proliferation, migration, and survival of cells [**113**, **114**, **115**].

There are two main types of RNA genes that are most often cited in relation to cancer: long non-coding RNAs (lncRNA) and microRNAs (miRNA) [**116**]. It should be noted that work has also been done on the role of circular non-coding RNA in cancer but these have been excluded from this thesis [**117**, **118**]. The presence of relevant lncRNAs and microRNAs

will be briefly discussed below.

Recent studies looking into lncRNA have found several genes that are implicated in the metastasis of breast cancers (*HOTAIR*), lung/cervical cancers (*MALAT1*), and prostate cancers (*PRNCR1* and *PCGEM1*) [**116**]. Examining the DeepLift results revealed that indeed *HOTAIR* was identified in metastatic BRCA and *PCGEM1* was identified in metastatic PRAD. Note that *MALAT1* was not found by the model within the metastatic BRCA cancer type. The model's results correlate with the literature and suggest that the model has, at least in part, the ability to detect useful biological insight from lncRNA genes, as well as coding genes. The inclusion of more lncRNA in the data set might provide a means for which to further expand the knowledge-base surrounding the role of lncRNA in various cancer types. Furthermore, the examples listed here are a subset of lncRNA genes available for study and simply show that this a feasible line of inquiry for further analysis.

With regards to microRNAs, there are a number that have been implicated across multiple cancers (discussed in Chapter 1) and have been shown to play a role in metastasis [**116**, **119**]. To reiterate, studies have found that miRNAs influence metastasis in a wide range of ways including by targeting oncogenes and/or tumour suppressors, modulating cancer stem cell properties, regulating EMT, and by influencing changes in the microenvironment. Furthermore, genes involved in the regulation of miRNA biogenesis have been implicated in cancer as well, adding an additional layer by which miRNAs themselves can be dysregulated [**119**]. Given the variety of ways in which miRNAs have been shown to influence metastasis (also see Chapter 1), the results obtained from the model seem to reflect this influence. With such a large number of cellular functions being affected by miRNA expression and such a high proportion of RNA gene utilization in the model, we can speculate that there are RNA expression patterns that can be learned to classify and potentially prognosticate metastatic cancers. In depth analysis of each cancer type's identified RNA genes and their functional annotations could be conducted to further glean biological insights.

Cancer Type Task Discussion: RNA Genes in Primary Cancer Types

It should be noted that all of the primary cancer types, with the exception of LAML, that are part of the organ systems that showed elevated RNA gene proportions (see the organ system of origin section above) have at least slightly elevated RNA gene proportions (above 0.03) at the cancer type level. The cancer types that are a part of the thoracic, hematologic, central nervous system, and soft tissue organ systems are as follows: CLL, CML, DLBC, FL, LUAD, LUSC, MB, MESO, and PCPG. We can speculate that some of the RNA genes identified as important within these cancer types are reflective of the organ systems from which they came.

Cancer Type Task Discussion: RNA Genes in Outlying Primary Cancer Types

There are three primary cancer types (Figure 4.16) that show high levels of RNA gene involvement consistent with the levels seen in metastatic cancers (between 23% and 37% RNA genes): CLL, FL, and MB. These RNA proportion values are high enough to be considered outliers from the rest of the primary cancer types. There are an additional 9 primary cancer types that have at least two times the RNA involvement when compared to the primary cancer types as a whole. These types can be seen in Figure 4.17. To validate the significance of these findings, there are two metrics that should be considered. The first is the classification performance (F1-score) and the second is the total number of important genes identified by the model. It is important to see if these cancer types differ from the other primary cancers in terms of either metric as the significance of the increase in RNA gene proportions may be related.

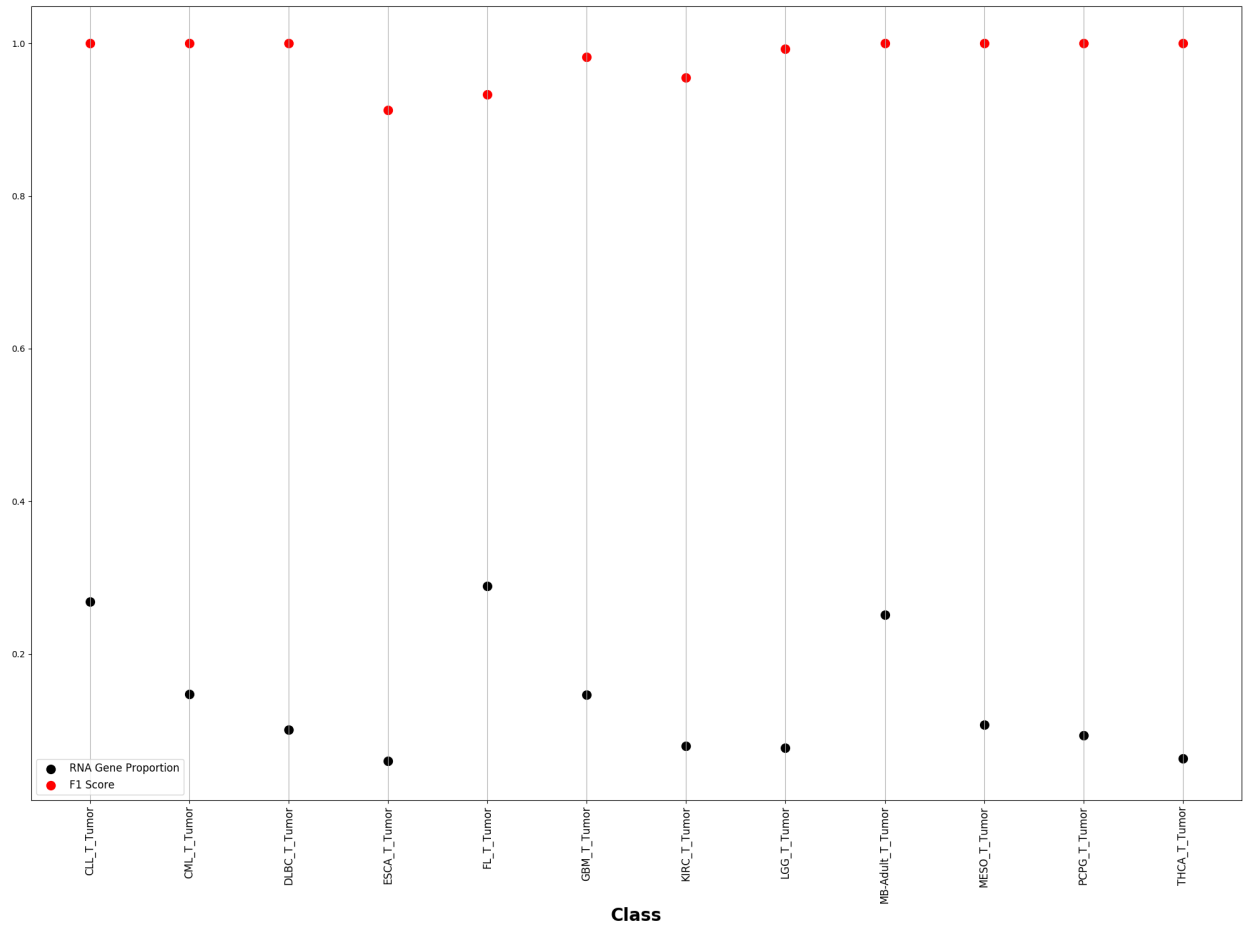

Figure 4.17: A scatter plot showing the proportion of RNA genes (black) within the positive important genes identified by DeepLift and the corresponding F1 classification scores (red) for primary cancer types whose proportions were greater 0.06

Figure 4.17 illustrates the RNA gene proportions in conjunction with the F1-scores for a subset of primary cancer types. The average RNA gene proportion across all primary cancer types is approximately 0.03. The subset of cancer types presented in Figure 4.17 was selected on the basis of having an RNA gene proportion of at least 0.06 (two times the average value for primary cancer types). The figure shows that the F1-scores for each of the primary cancer types listed remains close to 1.0. This seems to indicate that the RNA proportions are not a contributing factor on the classification performance. If this were the case, we would expect that cancer types with higher RNA gene usage would have poor performance relative to the

others.

To observe the impact of the total number of important genes on the RNA proportions reported, we need to look a bit more carefully at the results in Figure 4.16. Specifically, MESO and DLBC utilize nearly all of the available 26668 genes and as such, the proportion of RNA genes identified closely reflects the total number of RNA genes available within the data set (2890 genes or 10.8%). This effectively eliminates the significance of what appears to be higher RNA gene involvement for these two classes. As a result, we should exclude them from further analysis pertaining to RNA gene significance.

Having identified CLL, FL, and MB as primary cancer outliers, when we consider Figure 4.16, we note that they make up the primary cancer types with the lowest number of identified important genes, each with fewer than 10000. When compared to the bulk of the primary cancer types, which have approximately 18000 genes, this is a significant decrease. This suggests that perhaps the proportion of RNA genes is specific to these cancer types and may reflect the underlying biology. It also suggests that patterns of gene expression involving RNA genes are more easily learned by the model with the use of fewer genes. This also indicates that RNA gene expression patterns are more informative for classification than non-RNA genes.

The following sections will examine the biological underpinnings of RNA gene involvement for CLL, FL, and MB. As we will see below, each of these cancer types have literature to support that RNA genes, particularly miRNAs, play a significant role. It should be noted that miRNAs have been more widely studied and thus the literature and model results presented lean heavily towards miRNAs and away from other types of RNA genes (like long non-coding or circular).

Cancer Type Task Discussion: RNA Genes in Follicular Lymphoma (FL)

The involvement of RNA genes in lymphomas has been studied over the past decade and have been shown to have prognostic value [**120-127**]. Specifically, miRNA expression can be utilized to produce unique miRNA signatures that have indications with regards to treatment response for lymphomas [**120**, **121**, **122**, **123**]. A number miRNAs have been indicated in the development of follicular lymphomas through the regulation of *BCL2* with *miR-15* and *miR-16*, hematopoesis with *miR-150* and *miR-155*, and tumour development with *miR-210*, *miR-10a*, *miR-17-5P* and *miR-145* [**120**, **125**, **126**, **127**]. The genes identified by the model for FL contain examples from each of these regulatory categories and include *mir-15*, *miR-16*, *miR-150*, and *miR-210*. This illustrates the model's ability to learn some of the underlying miRNA signatures of FL. Note that there has been at least one long non-coding RNA gene (*RP11-625 L16.3*) identified as playing a pathogenic role in FL, but this gene was not present in the set of genes available to the model [**128**].

Cancer Type Task Discussion: RNA Genes in Chronic Lymphocytic Leukemia (CLL)

MicroRNA expression profiles have been shown to be of value in assessing the prognosis, progression, and drug resistance of CLL [**129**]. The following have been identified as the most deregulated miRNAs in CLL: *miR-15/16* cluster, *miR-34b/c*, *miR-29*, *miR-181b*, *miR-17/92*, *miR-150*, and *miR-155*. The model identified *miR-15b*, *miR16-1/2*, *miR-34b/c*, *miR-29b2*, *miR1-81b1/2*, and *miR-150* as being significant RNA genes in CLL [**128**, **129**]. These identified genes correspond well with the deregulated miRNAs from the literature on CLL and support the ability of the model to identify known, relevant RNA genes. Further analysis could seek to determine the expression levels of the relevant microRNA. Note that there exists some long non-coding RNA and circular RNA that may be implicated in CLL but none identified in the literature were found by the model [**128**].

Cancer Type Task Discussion: RNA Genes in Medulloblastoma (MB)

Recent studies have begun to establish the role of RNA genes in the development of medulloblastomas. One such study identified that MB can be differentiated from normal brain tissue using the expression profiles of the *miR-9* and *miR-125a* microRNA genes [**131**]. The model in this thesis supported the importance of these miRNAs and selected both of them for use in classifying MB. There are a number of other miRNAs that have been identified as either tumour suppressing or oncogenic and can be found in papers by Mollashahi et al. (2019), Cho et al. (2010), and Joshi et al. (2019) [**130**, **131**, **132**]. For example, Mollashahi et al. (2019) identified *miR-125b*, *miR-324-5p*, and *miR-32* as tumor suppressors within MB and indicated their dysregulation contributes to the development of MB [**130**]. Joshi et al. (2019) discuss the role of long non-coding RNA in MB and indicate that they are key regulators of cell proliferation and differentiation and that their dysregulation contributes to the development of many other cancers as well [**132**]. While data on lncRNA in MB is limited, the Joshi et al. (2019) paper lists 8 lncRNA genes implicated in MB. The model selected for 3 of the 8 genes and the results are presented in Table 4.9 below.

RNA Gene Type	**RNA Genes Found**
Oncomir	*miR-30b/d*, *miR-10b*, *miR-367*, *miR-106b*
Tumour Suppressor MicroRNA	*miR-193*, *miR-32*, *miR-124*, *miR-199b*, *miR-324*, *miR-326*, *miR-125a/b*, *miR-218*, *miR-31*, *miR-135a*, *miR-494*, *miR-221*
Long Non-Coding RNA	*CRNDE*, *LOXL1-AS1*, *NKX2-2AS1*

Table 4.9: List of RNA genes found by the model that are also implicated in Medulloblastoma

Cancer Type Task Discussion Summary: RNA Genes in Outlying Primary Cancers

Given the results presented in the preceding sections, it is clear that the model is learning something about the role of RNA genes within primary CLL, FL, and MB. It is encouraging

that the results span multiple types of RNA genes (lncRNA, oncomirs, and tumour suppressing miRNAs) and has overlap with genes identified in the relevant scientific literature. We may be able to identify key functions that are disrupted in each of the cancer types as a result of dysregulation within the identified RNA genes. It would also be interesting to compare and contrast the identified RNA genes with those found in the metastatic cancer types.

Further analysis of these outliers should look into the correlation between RNA type and expression levels to determine if the patterns match what is to be expected from a biological perspective. For example, we would expect to see oncomirs being overexpressed and tumour suppression miRNAs being underexpressed. This will require refining how the gene expression categories are defined.

Cancer Subtype Task Results

Figure 4.18 presents the proportion of RNA genes identified as important within each class of the cancer subtype task.

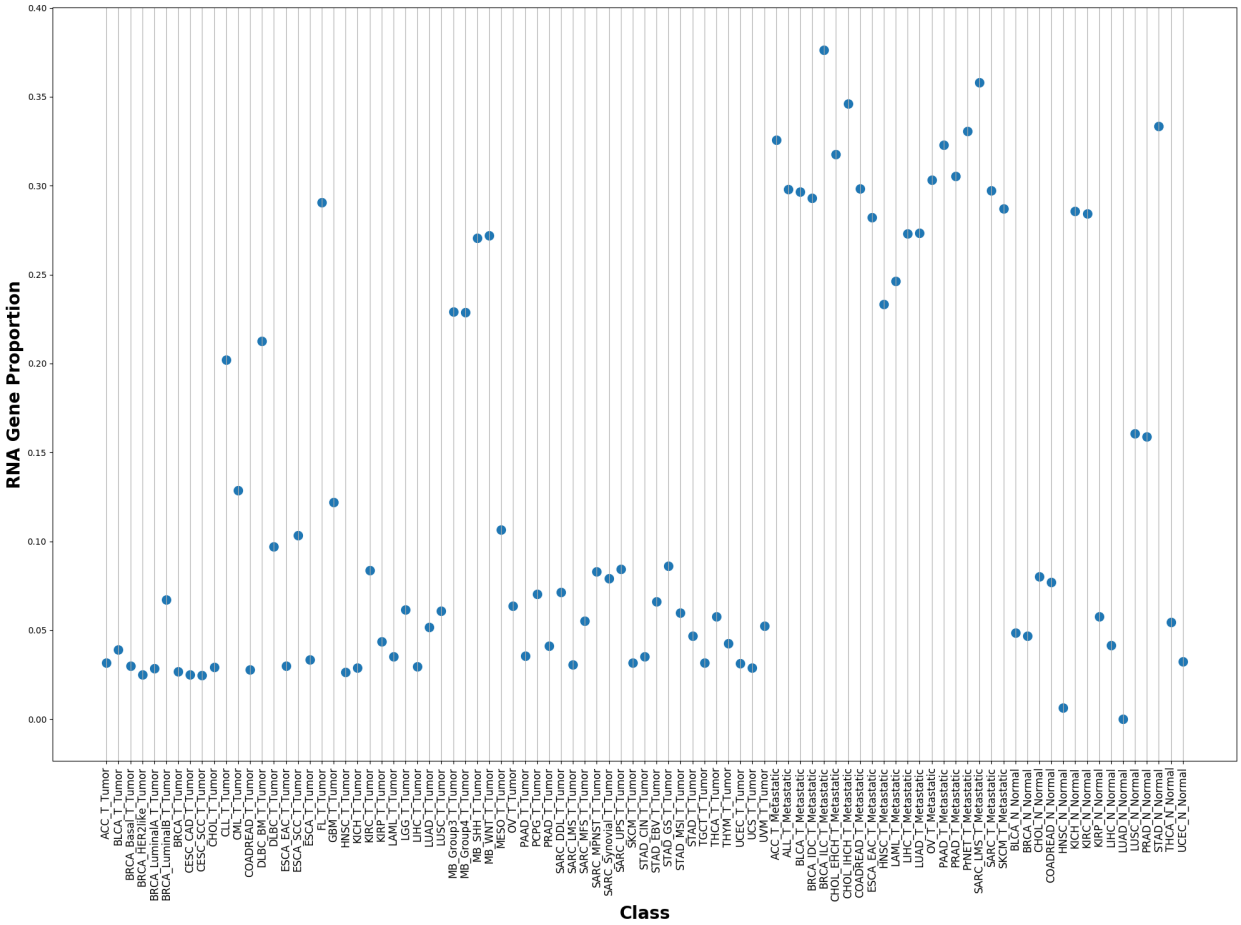

Figure 4.18: A scatter plot showing the proportion of RNA genes within the positive important genes identified for the cancer subtype classes.

In Figure 4.18, we see that the proportion of RNA genes utilized in the metastatic cancer subtypes range from 0.23 to 0.37 and are spread relatively evenly throughout this range. This is very similar to the results seen at the cancer type level as there are only 3 metastatic cancer types with subtype annotations.

The range for primary cancer types is 0.025 to 0.29 with the vast majority having an RNA proportion of approximately 0.03. When comparing the metastatic and primary subtype proportions, the majority of primary cancers again use 7 times fewer RNA genes than metastatic cancers. When comparing the cancer type and subtype proportions for primary

cancers, we note that the proportions have risen in approximately half of the subtypes making their values 0.05 or above. We also note that in addition to the three outlier cancer types (MB, CLL, and FL) seen in the previous section, at the subtype level, DLBC_BM should now be considered an outlier as well. DLBC_BM has an RNA gene proportion of 0.22, up from the DLBC proportion of 0.10 at the cancer type level.

The normal tissue classes remain identical to the proportions seen in the cancer type results.

Primary DLBC Bone Marrow Discussion

One notable oberservation of RNA gene involvement within cancer subtypes was the high RNA gene proportion for the primary DLBC_BM subtype. There is more than two times the RNA gene proportion in DLBC with bone marrow involvement (DLBC_BM) than in the DLBC subtype without. If we look more carefully at the number of genes involved (see Figure 4.6) we see that the number of genes utilized by DLBC is double that of DLBC_BM. So while DLBC_BM utilizes fewer genes, it retains the same number of RNA genes. This suggests two things about DLBC_BM. First, that it is easier to classify, as it requires fewer genes. Second, non-RNA genes were excluded in favour of RNA genes, implying an important role for RNA genes in bone marrow involvement of DLBC. According to the literature regarding DLBC, there are several miRNAs identified as being responsible for B cell development in bone marrow [**135**]. The first of these found was *miR-181a* and was indeed highlighted by the model as important [**135**]. The literature presents a thorough understanding of B cell development and how it correlates with miRNA expression. Further investigation into the miRNAs identified in the literature and those found within the DLBC subtypes by the model could provide further insight into the effect of differential expression on the progression of DLBC from a functional perspective. This would allow linking our understanding of normal and abnormal B cell development to the expression patterns of RNA genes.

Cancer Subtype Task Discussion

RNA Expression Summary

The results presented in the above sections attempt to validate the increased RNA gene importance seen in some primary and all metastatic cancer types. We have seen clear evidence that the model has uncovered significant differences in the contribution of RNA genes between metastatic cancers and the majority of primary ones. While the analysis given is far from exhaustive, it serves to show the potential for this line of research. We have identified a number of RNA genes across a variety of functions and cancer types that correlate, at least in part, with the relevant scientific literature. This has shown the capacity of the machine learning pipeline developed as part of this thesis to identify patterns in genes, coding or not, and that these genes may be biologically relevant.

There is an important caveat to consider when analyzing the RNA gene importance. All of the cancer classes (primary and metastatic) with the highest RNA gene proportions are classes in which the data has come from sources outside of the TCGA data set. Since TCGA data composes the bulk of the training data, it is possible that these results reflect some artifacts in the data that exist as a result of sequencing protocol (ie. batch effect). If this is the case, determining the true biological significance of increased RNA gene proportions requires careful further analysis of the results and the methods of data generation for each source. The implication of batch effect is discussed in more detail its own section below.

4.2.7 Pseudogenes

The results presented in this section serve to quantify the role that pseudogenes play in different cancers. The majority of the discussion that follows will, as with the RNA genes, focus on the disease state and cancer type classification tasks as the trends remain similar within the cancer subtype task. These classification tasks should sufficiently exemplify the

larger trends across cancers. It should be noted that there are 5280 pseudogenes in the full set of genes available to the model and that this comprises 19.8% of the available genes.

Organ System of Origin Task Results

Figure 4.19 presents the proportion of pseudogenes identified as important within each class of the organ system of origin task. We note that, as in the RNA gene results, the central nervous system, hematologic, and thoracic classes have the highest proportion of pseudogenes important for classification. We also note that these three classes are outliers from the other organ system classes with at least two and a half times the number of pseudogene involvement.

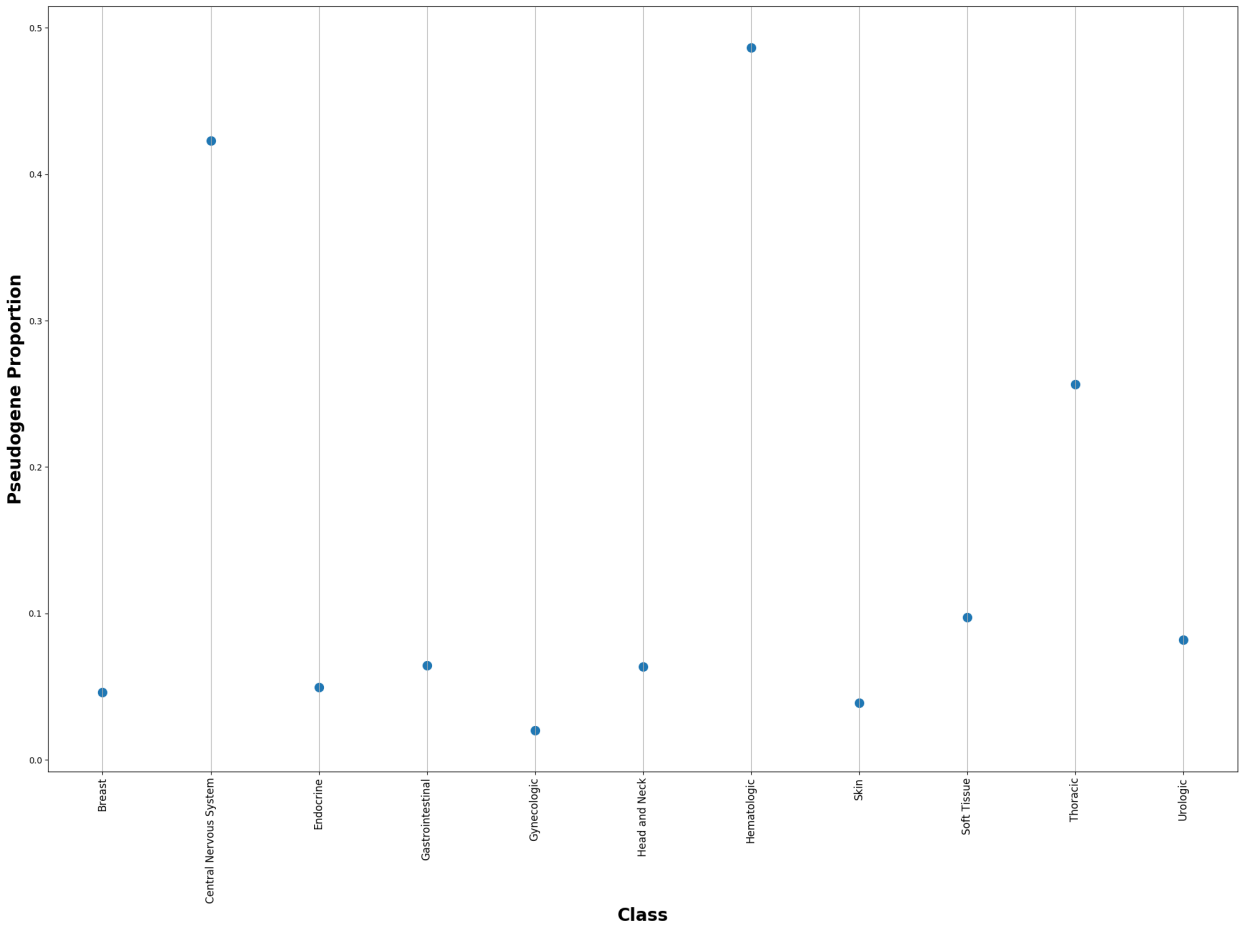

Figure 4.19: A scatter plot showing the proportion of pseudogenes within the positive important genes identified for the classes within the organ system of origin task.

Disease State Task Results

Figure 4.20 presents the proportion of pseudogenes identified as important within each class of the disease state task. We note again that, as in the RNA gene results, the metastatic class has the highest proportion. The metastatic class utilizes more than two times the number of pseudogenes in classification when compared to the primary class. Normal class classifications use very few pseudogenes.

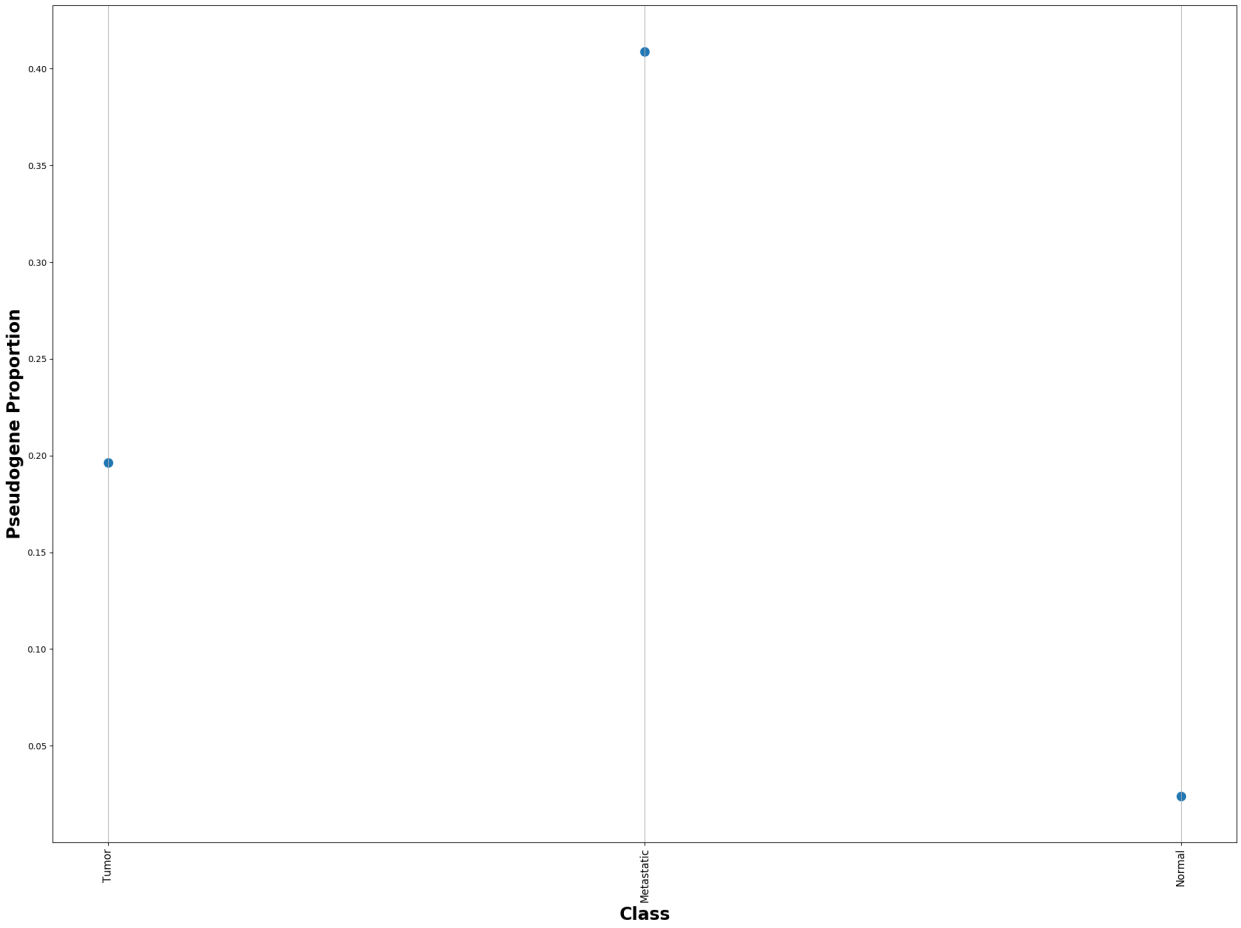

Figure 4.20: A scatter plot showing the proportion of pseudogenes within the positive important genes identified for the classes within the disease state task.

Cancer Type Task Results

Figure 4.21 presents the proportion of pseudogenes identified as important within each class of the cancer type task. We observe that the metastatic classes utilize significantly more pseudogenes than most primary and normal classes. The bulk of the metastatic classes have pseudogene proportions between 0.5 and 0.55. The majority of primary cancer types have a pseudogene proportion of approximately 0.045. There are some notable exceptions to the primary cancer types. The CLL, FL, and MB classes are outliers with pseudogene proportions of 0.55, 0.585, and 0.48 respectively. There are another 11 primary cancer types that show at least two times the number of pseudogenes as the majority. These primary cancer types are as follows: CML, DLBC, ESCA, GBM, KIRC, LGG, LUAD, LUSC, MESO, PCPG, THCA, and THYM.

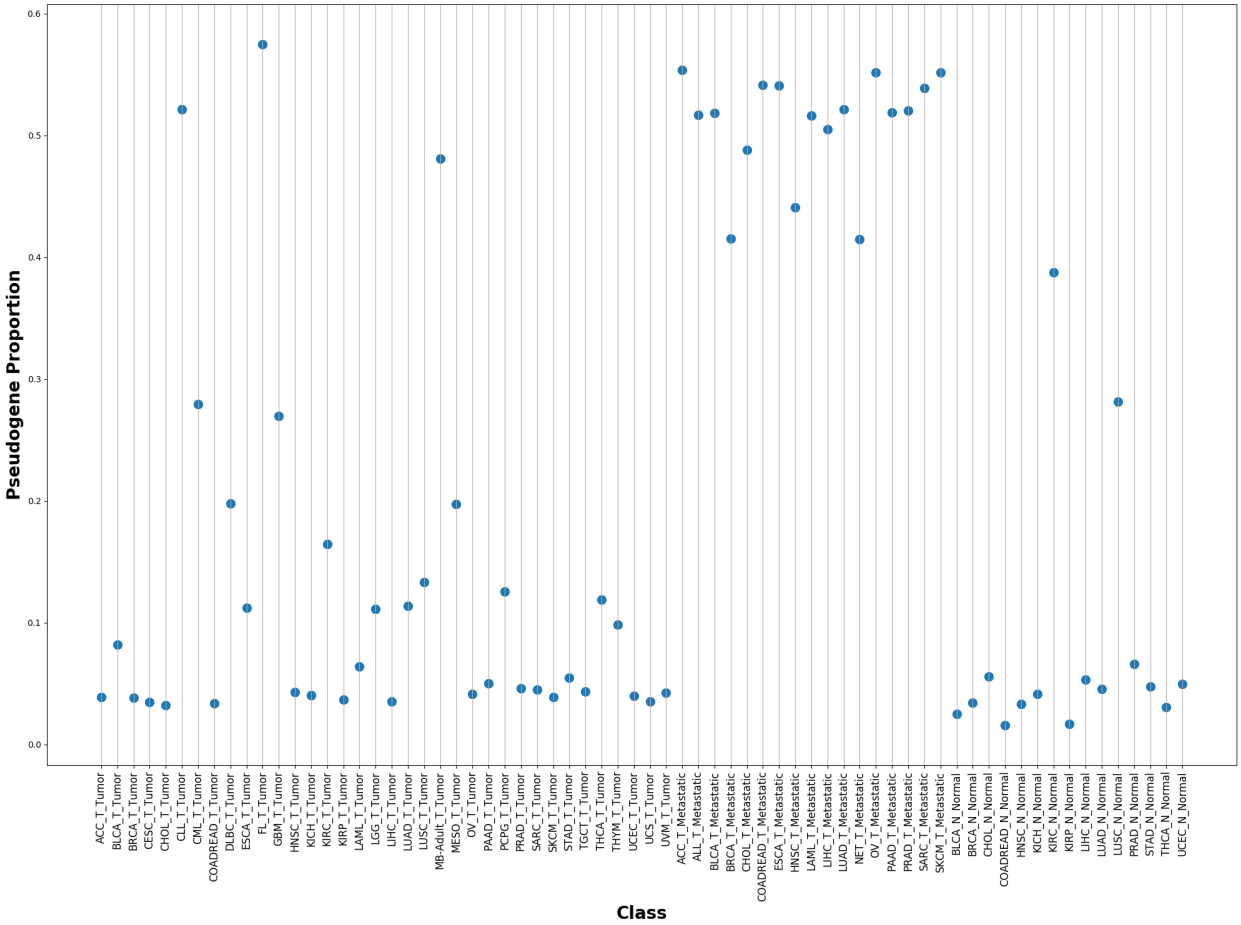

Figure 4.21: A scatter plot showing the proportion of pseudogenes within the positive important genes identified for the classes within the cancer type classes.

Cancer Type Task Discussion

The results presented here show that pseudogenes play a large role in classifying metastatic cancer. In fact, at or near 50% of a classification decision is made using pseudogenes for all metastatic cancers. This value exceeds that of the RNA gene proportions seen in the previous section. Studies have suggested that the diagnostic and prognostic power of pseudogenes is in some cases higher than that of miRNAs [**136**]. They have also found that there are specific signatures of pseudogenes that correlate with poor survival and as a result could suggest a predisposition to metastasis [**136**, **137**]. The implications of these studies support the results presented above for all of the metastatic cancers. In other words, the model has

placed a high importance on pseudogenes for differentiating metastatic from primary cancers and may prove useful in future work for determining a predisposition for metastasis.

Similar to the caveat placed on the RNA gene results, we must consider the implication of the data source on the pseudogene content of cancer type classifications. We again see that the metastatic proportions and the proportions for the largest primary cancer outliers (CLL, FL, and MB) are all cancer types that come from non-TCGA datasets. The implication of this is that there may be some batch effect occurring. This consideration somewhat diminishes the reliability of the trends shown in these results and requires further investigation to alleviate. However, when we examine the batch effect (see the relevant section) based on the data sources, we note that the metastatic cancers appear to be the most problematic. We also note that there are 11 other primary cancers (see above) that show at least two times the pseudogene proportions of the majority of primary cancers. Of these 11 cancer types, eight of them were sourced from TCGA. This suggests that while batch effect may play a role in pseudogene importance, there are a number of examples where this is not the case and patterns of pseudogene expression may be strictly a result of underlying biological. Further in depth analysis of the genes identified for the eight TCGA cancer types with high pseudogene importance could serve to better support the biological relevance of pseudogene expression.

Cancer Subtype Task Results

Figure 4.22 presents the proportion of pseudogenes identified as important within each class of the cancer subtype task. Generally, the observed trends seen here are very similar to those found within the cancer type task. The metastatic classes all have significantly higher pseudogene proportions than the bulk of the primary ones, with values ranging from 0.405 to 0.57. The outlier primary cancer subtypes correspond to the outlier primary cancer types, with CLL, FL, and MB subtypes composing this group. There are, however, more classes that show higher levels (above 0.045) of pseudogene involvement than we saw at the cancer

type level. We note that 24 of the 55 primary cancer subtypes have values close to 0.045 and the range of value when compared to the cancer type results is much more varied. This means the majority of subtypes show some elevated levels of pseudogene involvement compared to their corresponding cancer types. The significance of this is that some subtypes show elevated pseudogene involvement when compared to their related subtypes (ie. within the same type). For example, Luminal B BRCA has 3 times higher pseudogene involvement than the other BRCA subtypes and this elevated pseudogene involvement was not visible at the cancer type level. We again see this in the STAD subtypes with STAD_EBV and STAD_MSI being elevated and in ESCA with ESCA_SCC being elevated.

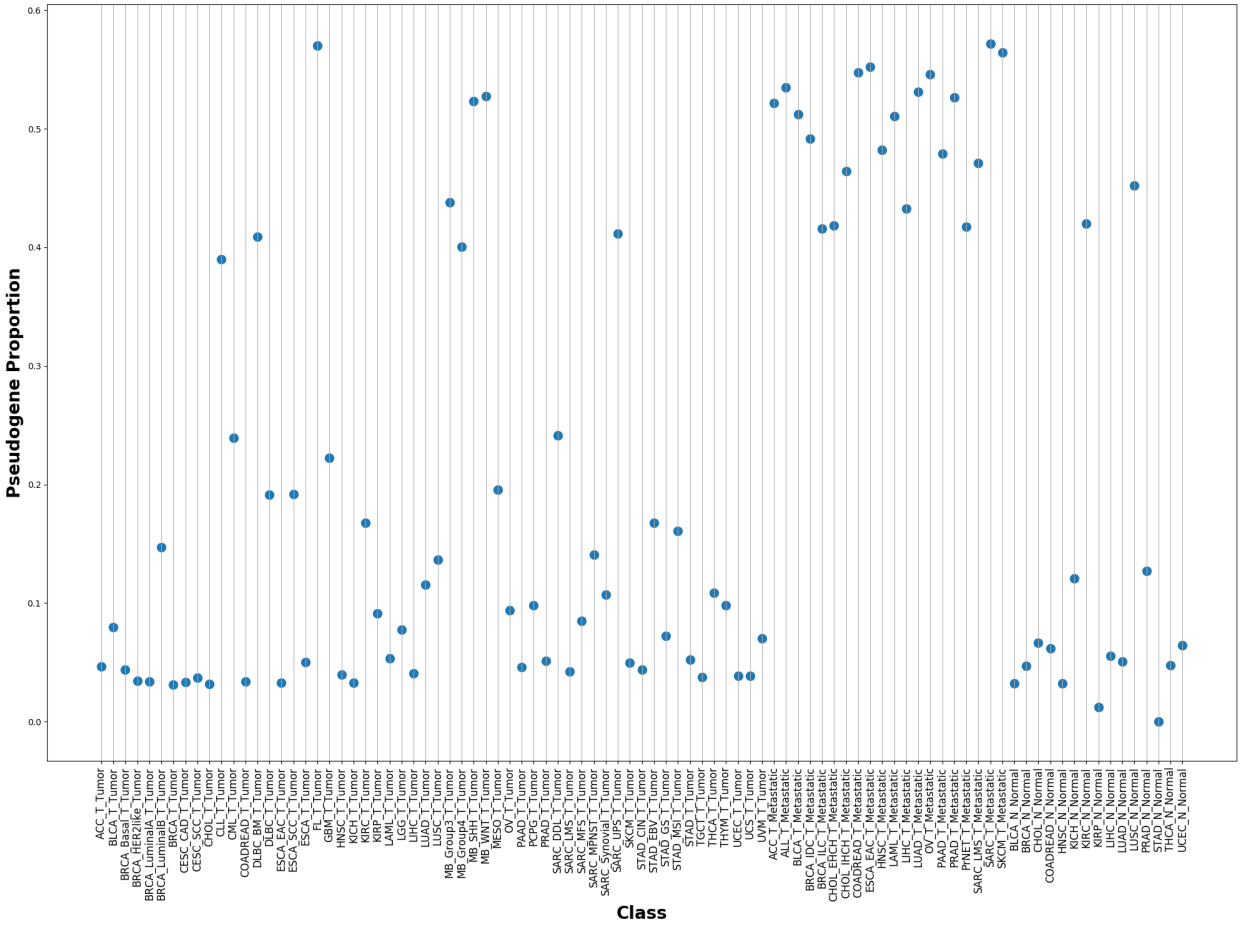

Figure 4.22: A scatter plot showing the proportion of pseudogenes within the positive important genes identified for the classes within the cancer subtype task.

Cancer Subtype Task Discussion

We noted similar trends in the cancer subtype pseudogene proportions to those seen at the cancer type level. The implication for metastatic cancers is still that pseudogenes play an important role in their classification. We again have the caveat that the influence of data sources should be considered because some of the subtypes with elevated pseudogene importance are from non-TCGA data sources.

The most noteworthy change seen in the results from cancer type to subtype pertains to the elevated pseudogene levels of one or more subtypes where the other related subtype proportions remain low. One clear example of this is with the primary Luminal B BRCA subtype. Luminal B BRCA shows 3 times the level of pseudogene usage when compared to the other primary BRCA subtypes. One study has supported the use of pseudogenes for discerning breast cancers from normal tissue and other breast cancer subtypes [**138**]. This study suggests that pseudogenes are a valid line of inquiry for discerning breast cancer subtypes [**138**]. Given that the primary BRCA subtypes all comes from the same data set and that there is a marked change in one subtype compared to both the related cancer type and subtypes, this is a strongest indication for a biological interpretation of the impact of pseudogenes. Couple these facts with supporting evidence of the value of pseudogenes in cancer diagnosis from the literature, and pseudogene analysis seems to be a viable avenue for which to further characterize and diagnose cancers. Continued analysis of the role of pseudogenes on cancer diagnosis and the characterization of cancer subtypes should focus on the kinds of examples where individual subtypes differ from their related subtypes in order to remove batch effect implications.

4.3 The Implications of Batch Effect

This section will present and discuss the implications of batch effect on the biological interpretation of the above results. Figure 4.23 presents a t-SNE of the transcriptome data

from the full training data set.

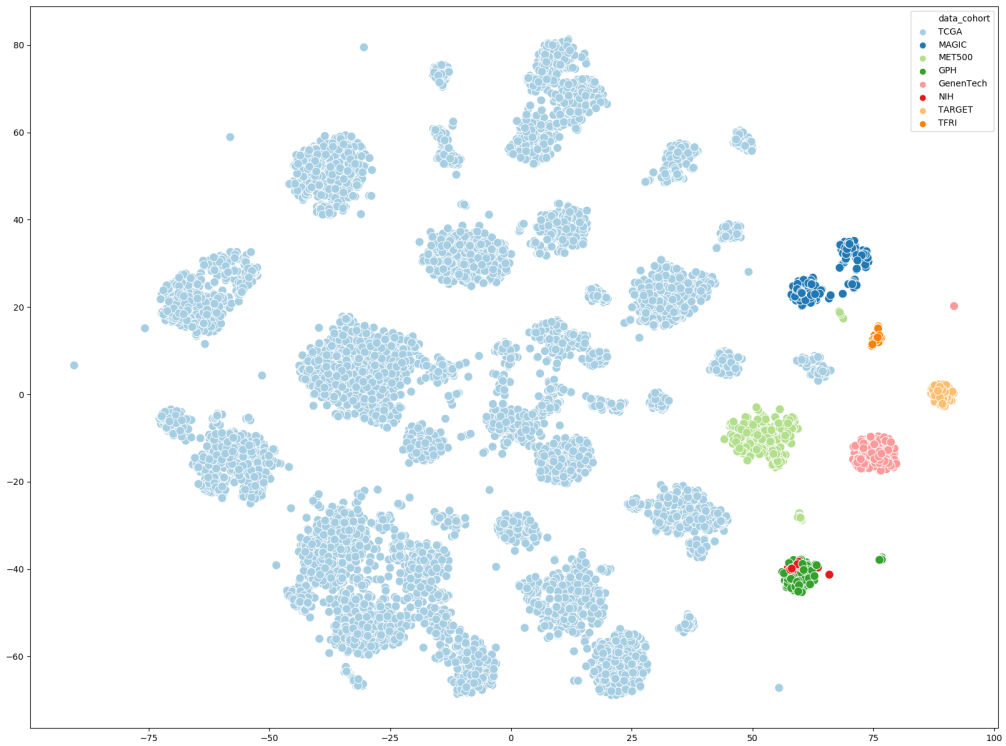

Figure 4.23: A t-SNE plot of the transcriptome data for the full training data set coloured by data cohort.

Given the results shown by Figure 4.23, there appears to be a potential issue with batch effect within the training data set. In order to understand the implication of possible batch effect, we must look at the cancer types within each data cohort. Table 4.10 shows the relevant data.

Data Cohort	Cancer Type
GPH	CLL_T_Tumor & DLBC_T_Tumor
NIH	FL_T_Tumor & DLBC_T_Tumor
MET500	All Metastatic Cancer Types
GenenTech	MESO_T_Tumor
TARGET	CML_T_Tumor
TFRI	GBM_T_Tumor
MAGIC	MB-Adult_T_Tumor

Table 4.10: List of cancer types and the non-TCGA data cohorts from which they came.

4.3.1 Batch Effect Implications on the Interpretation of Metastatic Cancers

When we consider Table 4.10, we see that batch effect is potentially a serious problem for the metastatic cancers (MET500). All of the cancer types group together based on data cohort. This observation has a negative implication on the ability to interpret any biological features of metastatic cancers within the disease state task. Within this task, the model would likely be able to accurately classify metastatic cancers on the basis of features present within the data source alone. The interpretation of metastatic cancer results within the disease state classification task should consider these effects carefully.

The batch effect is such that the shared trend of high RNA gene and pseudogene importance across all metastatic cancers may not be a reliable source for biological interpretation. The implication is not that all of the important genes found for each metastatic cancer type are biologically irrelevant, but rather that a portion of them may be shared among all metastatic types and result as an artifact of the data set. The advantage to having multiple learning tasks, however, is that we are able to follow the important genes from the disease state (covering all metastatic cancers at once) down into the cancer type and subtype levels

and potentially filter out common features whose presence may exist due to batch effect. Also, since the model is encouraged to learn disease state features upstream of cancer type and subtype classifications, the important genes within the cancer type and subtype classes are genes used to differentiate between not only metastatic and primary cancers as a group, but also between individual metastatic cancers. This means that for interpretation purposes we can still expect the model to learn some unique features of each metastatic cancer type and subtype. However, we must be careful in considering overall trends seen across all metastatic cancers because as a whole, some of their important genes will be artifacts of the data source from which they came.

To support the conclusion that some of the biological features of metastatic cancers can be learned at the cancer type level, we can visualize the metastatic samples separately from the primary ones. Figure 4.24 is a t-SNE plot of only the metastatic samples from the training data and is coloured according to cancer type. This figure illustrates that there is at least some grouping together of cancer types in the metastatic domain. This suggests that there may be some common features for the model to learn within each cancer type and that the data is not entirely useless. Figure 4.25 is a t-SNE plot of only the primary cancer samples from the training data and is also coloured according to cancer type. Comparing Figures 4.24 and 4.25, we see that the metastatic cancer types are much less well defined and have greater overlap between types than the primary cancer types seen in Figure 4.25. It is possible that if we had more metastatic samples from each cancer type we would have better established groupings within the metastatic cancers. Given the results shown here with the current data set, we would expect worse classification performance on the metastatic cancers than the primary ones. This is, in fact, what is observed throughout this thesis and can partially be explained by batch effect and the feature set contained within the data.

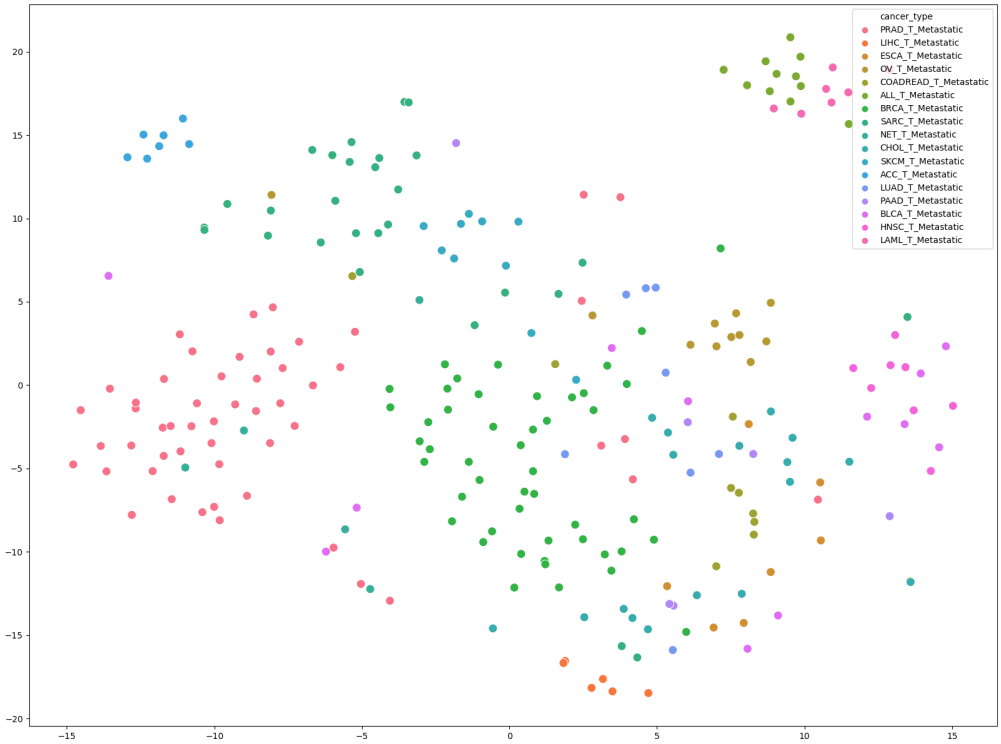

Figure 4.24: A t-SNE plot of the transcriptome data for the metastatic cancer types from the training data set coloured by cancer type.

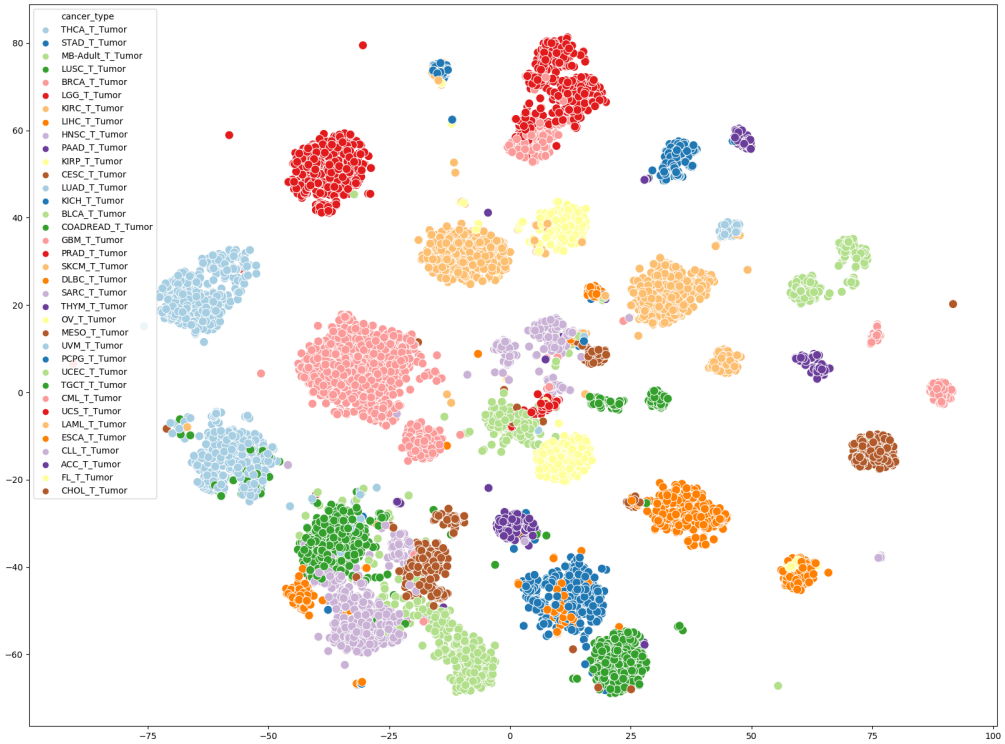

Figure 4.25: A t-SNE plot of the transcriptome data for the primary cancer types from the training data set coloured by cancer type.

4.3.2 Batch Effect Implications on the Interpretation of Primary Cancers

For the primary cancers from TCGA-external data sources, we must also consider the possible implications of batch effect. The cancer types listed in Table 4.11, with the exception of DLBC_T_Tumor, are all predisposed to suffering from batch effect. The reason for this is that they are each sourced from single data sources that are unique to their cancer type. This offers the machine learning model an opportunity to identify and leverage features present in the data stemming from the source as opposed to features relevant to the underlying biology.

Looking at Figure 4.26, we can exclude DLBC_T_Tumor from batch effect as it has multiple data sources in which to encourage the model to identify biologically representative features.

Data Cohort	**Cancer Subtype**
TCGA	DLBC_T_Tumor
GPH	DLBC_T_Tumor
NIH	DLBC_BM_T_Tumor

Table 4.11: List of DLBC cancer subtypes and the data cohorts from which they came.

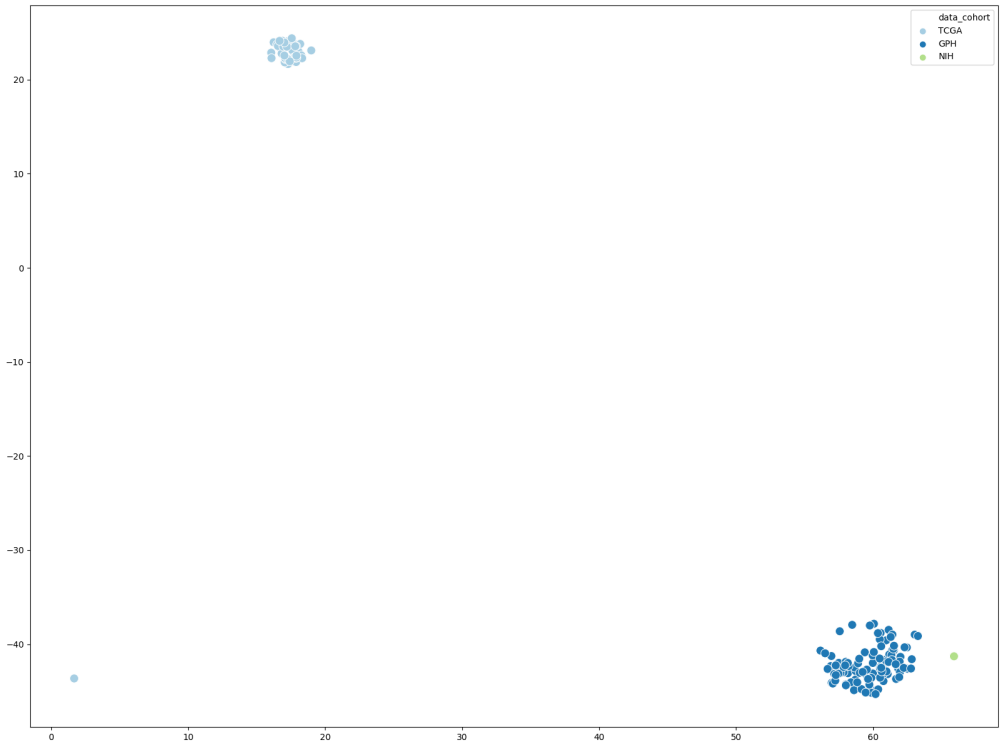

Figure 4.26: A t-SNE plot of the transcriptome data for the DLBC cancer type coloured by data cohort.

We can also identify a correlation between single data sources for primary cancers and the outlier types identified within the RNA gene and pseudogene results. CLL, FL, and MB each have a single data source and may have elevated importance of these gene types as a result of their data source. In light of this, future analysis of the RNA gene and pseudogene trends seen within primary cancer types should focus on those cancer types that do not appear to suffer from batch effect (ie. those from within TCGA). There were 11 other cancer types that showed elevated RNA gene and/or pseudogene importance, with eight of them being from within the TCGA data set.

4.3.3 Batch Effect Conclusion

The overall conclusion is that batch effect is potentially a problem within this data set. The nature of having limited and unique data sources for some cancer types has implications on the ability to interpret the biological implications of the results presented in this thesis. There are techniques that could be applied to the training data to try and mitigate these effects. For example, *ComBat-seq* is a recently published (January, 2020) tool for mitigating batch effect on RNA-seq data [**139**]. The model used for this thesis could be trained on batch corrected data and the results re-evaluated. Barring batch correction and retraining, the multi-task nature of the model provides a mechanism by which the filtration of gene results can be conducted and the significance of genes biologically interpreted. This could be done by leveraging the features identified within the disease state against the cancer type and subtype task results. Finally, the inclusion of at least two data sources for each cancer of interest could help to mitigate batch effect by encouraging the model to find common features between the sets.

The feature set (genes) is another aspect to consider as part of the batch effect analysis. The feature set for this thesis work contains the intersection of genes that are found within all of the data cohorts combined. Given the results presented in Figure 4.24 and 4.25, we can see that primary cancers types are better defined by the present feature set when compared

to the metastatic types. The metastatic types may need to include a different set of genes in order to be better differentiated from each other. By restricting the set of genes for the data set to the intersecting genes from each data source, we are almost certainly disposing potentially useful information. It may be the case that the metastatic cancers suffered a greater loss of information than the primary ones as a result of the gene exclusion conducted to generate the data set for this thesis.

4.4 Summary

The results presented in Chapter 4 were given in five sections. First, the biological validity of the gene results were examined using the normal tissue classes. The list of genes for some normal classes were functionally annotated using DAVID to identify enriched pathways that are indicative of the expected biological functions of the classes in question. Thyroid tissue showed enrichment of pathways involved in neuropeptide hormone production, lung tissue showed enrichment of gas exchange and oxygen binding/transport pathways, and breast tissue showed enrichment of pathways involved in milk and keratin production.

Following the biological validation of normal tissues, the number of important genes identified by the model for each class was presented. We noted that the number of important genes used for classification was significantly smaller for the metastatic and normal classes than for the primary ones. These results suggest that metastatic and normal tissues had more distinct patterns of expression and thus required fewer genes to identify. These results also suggest that as a whole, metastatic cancers have unique expression signatures that differentiate them from primary cancers. Whether or not this is biologically significant will require addressing the batch effect noted in the previous section.

The expression levels of the identified important genes was also evaluated. We noted that the model favoured genes with high expression levels in metastatic and normal classes within

the disease state task. We again saw this trend at the cancer type and subtype level. We concluded here that further analysis of the implication of the expression levels of important genes would require redefining of the over and underexpression categories on a per-class basis in order to confirm the biological impact of expression levels. The only conclusion that can be made from these results is that the model seems to need fewer genes for classification when those genes have expression levels far above the mean (over two standard deviations). Whether this is the result of biological or computational factors remains to be determined.

After identifying the number of important genes and their expression levels, some insight into the functionality of the important genes for the non-normal classes were examined. DAVID was again used for the functional annotation of enriched pathways. Here we correlated the enrichment of particular functional pathways (the top 10 most enriched) within metastatic and primary classes of interest with scientific literature that supports the function and presence of these pathways within each class.

Significant enrichment of microRNAs were found in the enriched pathways of the metastatic class within the disease state task. This prompted further analysis of the RNA gene content of each classes' important genes. The role of RNA genes was quantified and presented. We noted that there was a significant increase in the number of RNA genes identified as important within all metastatic cancers and three primary cancer types (CLL, FL, and MB). The role of RNA genes in each specific primary cancer type was discussed and shown to have support from the relevant scientific literature. The role of RNA genes in metastatic cancers was also discussed and the foundation for future research in this area was laid. The overall conclusion was that the model elected to make approximately 30% of a metastatic cancer classifications using RNA genes. This suggests that RNA genes play a large enough role in metastatic cancers when compared to primary ones to elicit a recognizable pattern of expression. This pattern can be effectively leveraged by a machine learning model and further analysis of the genes involved could result in novel insights on the progression of metastatic cancers. We noted the caveat that sequencing protocol and data generation may

have impacted the apparent important role of RNA genes in classification (ie. batch effect). While this caveat does not negate the presence of the particular RNA genes noted in the discussion, it could impact the perceived strength of the correlation between cancer type and the number of RNA genes present in the important genes listed for each class.

Finally, the role of pseudogenes in classification was presented. We noted a similar trend to the one found in RNA genes. The model elected to identify metastatic and normal classes with a high proportion of pseudogenes relative to both coding and RNA genes. This suggests that there is an expression pattern within pseudogenes that is of higher importance in classifying metastatic cancers than in primary ones. However, as with RNA gene importance, the biological significance of this needs to be further elucidated while carefully considering the implications of batch effect.

These five ways of examining the model's results have provided some examples of the kinds of data contained within the model's output. The value of these results lies in the large amount of data being output and that it appears to have some biological significance. Further study of the model's output should provide a means with which to gain insight into the biological functionality of genes within and across a variety of both metastatic and primary cancers.

Chapter 5

Conclusion

5.1 Summary of Findings

The first goal of this thesis research was to demonstrate the ability to classify with reasonable accuracy a set of normal, primary cancer, and metastatic cancer samples using gene expression data. Chapter 2 presented detailed information on the data set utilized for this work along with the methodology used to generate the machine learning model to be used for this task. This chapter also presented the results of model validation using five-fold cross-validation and a set of multi-task models with different combinations of learning tasks including organ system of origin, disease state, cancer type, and cancer subtype. We noted that the performance was relatively similar between the different models and since the eventual goal of this thesis was to produce and analyze the largest amount of data with as much granularity as possible, a multi-task model (referred to in Chapter 2 as the "all task" model) that included all four listed learning tasks was deemed appropriate for further use.

Having validated the model architecture and set of learning tasks in Chapter 2, Chapter 3 was focused on presenting the results of classification using a model trained on the full training data set. The classification performance was evaluated across each of the learning

tasks using two test sets. The test sets included one held-out data set composed of normal, primary cancer, and metastatic cancer samples, and an external test set (POG) composed of only metastatic samples. The overall trend was such that as the learning task increased in complexity and biological granularity (ie. from organ system to cancer subtype), the classification performance declined on both test sets. The model also performed worse on metastatic cancer classification than on primary cancer. This can be, at least in part, explained by batch effect. We noted that a t-SNE plot of the metastatic cancers does not differentiate as well into cancer type as primary cancers do. In addition, the training data set contains a large class imbalance that most certainly contributes to the reduction in performance on the metastatic classes, as they are in the minority. Furthermore, as the learning tasks increase in biological granularity (ie. from organ system to cancer subtype) the class sizes decrease by virtue of now having more classes with a data set that remains the same. Overall, while there were some exceptions noted, the majority of classes within each learning task were classified reasonably well. We determined that the classification performance was sufficient to warrant further downstream analysis of the model using DeepLift.

Chapter 4 of this thesis focused on extracting biological information from the trained multi-task model in an attempt to glean insight about the characteristics of various cancers. This task was accomplished using a backpropagation-based tool called DeepLift. DeepLift is designed to query the impact of input features on the output of a trained neural network. In the context of this thesis, it provided a mechanism to query what importance each gene plays on the classification of each class within each learning task. Using DeepLift, we were able to score the importance of each gene on the classification of cancers and use these scores to examine five aspects of the results.

The first aspect was to examine the number of important genes used for the positive classification of each class and it was determined that metastatic cancers had many fewer genes involved. This suggests that they have a unique pattern of expression that easily differentiates them from primary cancers. This may simply be the result of batch effect,

particularly within the disease state task.

The second aspect of the DeepLift results examined were the expression values of important genes selected by the model. We noted that when differentiating metastatic cancers within the disease state level the model selected to use only overexpressed genes. The caveat of this observation was that the definition of mean gene expression was skewed towards primary cancers and should be redefined to further refine the analysis of these results.

The third aspect of the DeepLift results investigated were the functional annotations of enriched pathways for some classes of interest. Three normal tissue types were used to validate the results of the functional annotations and the genes selected by the model. Following this, the enriched pathways within the primary and metastatic classes from the disease state task were presented and discussed.

The fourth analysis of the results looked at RNA gene importance for each class across every learning task. The results indicated that all metastatic and three outlier primary cancer types (CLL, FL, and MB) had significantly increased RNA importance assigned by the model when compared to the bulk of the primary and normal classes. A literature search was conducted to try and validate some of the RNA genes identified for metastatic cancers in the disease state task and the three outlier primary types in the cancer type task outputs. We noted that the model identified a number of microRNA genes within each of the cancers investigated and had literature supporting their role in each type. We did, however, observe that there was an apparent correlation between increased RNA gene importance and classes whose data came from non-TCGA sources. The implication of this is that there is some batch effect going on. These results should be further examined to determine if the model is truly learning biologically relevant trends in RNA gene involvement or simply learning to differentiate something in the sequencing process that is representative of the source of the data.

Finally, the results were examined for pseudogene importance in each class across each learning task. As with the RNA genes, we observed that all metastatic and three outlier primary cancer types (CLL, FL, and MB) had significantly increased pseudogene importance when compared to the bulk of the primary and normal classes. We again saw a correlation between non-TCGA data sources and an increase in pseudogene usage. This trend again raises concerns about batch effect. However, we also noted that there were 11 cancer types (excluding CLL, FL, and MB) that showed at least two times the pseudogene importance of most of the primary cancers. Within these 11 cancer types, there were eight that came from the TCGA data set and thus would be a good place to begin further analysis into the value of pseudogene characteristics. By focusing on these eight TCGA-sourced cancer types, we could get around some of the negative implications of batch effect and be more comfortable in making a biological interpretation of pseudogene and RNA gene trends.

5.2 Future Work

There are many avenues available for future work related to aspects of this thesis work. The first avenue to explore would be trying to batch correct the data. *ComBat-seq* was released in January of 2020 and may prove to be a useful tool for correcting the training data [**145**]. Following this, we should iterate on and improve the model used for classification. There may be changes to hyperparameters or architecture that could provide improved performance given the batch corrected data.

Following this, the way in which learning tasks interact within the model could be explored. For example, each non-terminal learning task's prediction could be included as an input feature to the downstream tasks' corresponding layers. The effect of this would be such that within each learning task, the model would be informed of the previous task's classification. This may help to mitigate compounded errors caused by having kept each learning tasks prediction separate. In other words, the model would have the ability to correct upstream learning task classification errors in downstream tasks by learning the relationship between

accuracy and the previous task's prediction.

Finally, the utility of the learning tasks could be further interrogated. For example, we saw that in some cases, including all learning tasks performed worse than a model missing the cancer type task. Depending on the intended use and desired output of the DeepLift data, we may benefit from excluding particular learning tasks and obtaining better accuracy on fewer tasks.

The bulk of the DeepLift results are left to be further investigated. There is an opportunity to conduct an in-depth review of the important genes for each cancer type and subtype. This thesis noted some larger trends that separate primary and metastatic cancers, but these results may be confounded by batch effect. Regardless, there remains numerous lists of genes available for each cancer type and subtype in which the negative implications of batch effect should be mitigated. Functional annotation of the important genes for each cancer type and subtype can and should be investigated for existing and novel pathways that may prove to be cancer driving.

Another avenue of research may be to try and differentiate metastatic and primary cancers at the disease state level using subsets of the gene types. We could attempt to classify the same cancers using only coding genes or only non-coding genes and compare the results to gain a more granular understanding of the impact of each gene type. Given that the genes used for each metastatic cancer classification were at least 24% RNA genes and at least 40% pseudogenes, there is reason to believe these types of genes encode a significant amount of information that can be used to differentiate these cancers. However, by focusing on these types of genes independently, we would be encouraging the model to learn more complex expression patterns within these genes. We may also be able to use these new results to better observe the implication of batch effect on the current set of results

Finally, for post-classification analysis it may be valuable to separate the primary and metastatic cancer classifications into separate models. This would encourage the model to learn more unique features of each cancer type as the broad, significant differences seen between primary and metastatic cancers, such as RNA and pseudogene expression, could not be as easily leveraged. In the current state, the model is able to use broad categories of genes such as pseudogenes to make the bulk of a classification decision for metastatic cancers, as it is vastly different from the primary ones. It would be interesting to see the impact on the results when the model is forced to choose between cancer types that are more closely related as far as non-coding genes are concerned. This would also mitigate issues related to batch effect.

Bibliography

[1] Brücher, B. L., Jamall, I. S. (2014). Epistemology of the origin of cancer: a new paradigm. BMC cancer, 14(1), 1-15.

[2] Brenner, D. R., Weir, H. K., Demers, A. A., Ellison, L. F., Louzado, C., Shaw, A., ... Smith, L. M. (2020). Projected estimates of cancer in Canada in 2020. Cmaj, 192(9), E199-E205.

[3] Government of Canada, Statistics Canada. (2020, January 29). Number and rates of new cases of primary cancer, by cancer type, age group and sex. https://www150.statcan.gc.ca/t1/tbl1/en/tv.action?pid=1310011101.

[4] Seyfried, T. N., Flores, R. E., Poff, A. M., D'Agostino, D. P. (2014). Cancer as a metabolic disease: Implications for novel therapeutics. Carcinogenesis (New York), 35(3), 515-527. doi:10.1093/carcin/bgt480

[5] Dillekås, H., Rogers, M. S., Straume, O. (2019). Are 90% of deaths from cancer caused by metastases?. Cancer medicine, 8(12), 5574–5576. https://doi.org/10.1002/cam4.2474

[6] Chaffer, C. L., Weinberg, R. A. (2011). A perspective on cancer cell metastasis. Science (American Association for the Advancement of Science), 331(6024), 1559-1564. doi:10.1126/science.1203543

[7] Martin, T. A., Ye, L., Sanders, A. J., Lane, J., Jiang, W. G. (2013). Cancer invasion and metastasis: molecular and cellular perspective. In Madame Curie Bioscience Database [Internet]. Landes Bioscience.

[8] Staub, E., Buhr, H. J., Gröne, J. (2010). Predicting the site of origin of tumors by a gene expression signature derived from normal tissues. Oncogene, 29(31), 4485-4492.

[9] Hess, K. R., Varadhachary, G. R., Taylor, S. H., Wei, W., Raber, M. N., Lenzi, R., Abbruzzese, J. L. (2006). Metastatic patterns in adenocarcinoma. Cancer, 106(7), 1624-1633.

[10] Anderson, G. G., Weiss, L. M. (2010). Determining tissue of origin for metastatic cancers: meta-analysis and literature review of immunohistochemistry performance. Applied immunohistochemistry molecular morphology : AIMM, 18(1), 3–8.

[11] Brosius, J. (2009). The fragmented gene. Annals of the New York Academy of Sciences, 1178(1), 186-193.

[12] Brown, T. A. (2018). genomes 4. CRC Press. https://doi.org/10.1201/9781315226828.

[13] Shin, B. K., Wang, H., Yim, A. M., Naour, F. L., Brichory, F., Jang, J. H., Zhao, R., Puravs, E., Tra, J., Michael, C. W., Misek, D. E., Hanash, S. M. (2003). Global profiling of the cell surface proteome of cancer cells uncovers an abundance of proteins with chaperone function. The Journal of Biological Chemistry, 278(9), 7607-7616. https://doi.org/10.1074/jbc.M210455200.

[14] Zhang, L., Zhou, W., Velculescu, V. E., Kern, S. E., Hruban, R. H., Hamilton, S. R., ... Kinzler, K. W. (1997). Gene expression profiles in normal and cancer cells. Science, 276(5316), 1268-1272.

[15] Golub, T. R., Slonim, D. K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J. P., ... Bloomfield, C. D. (1999). Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. science, 286(5439), 531-537.

[16] Croce C. M. (2009). Causes and consequences of microRNA dysregulation in cancer. Nature reviews. Genetics, 10(10), 704–714. https://doi.org/10.1038/nrg2634

[17] Shin, D. M., Kim, J., Ro, J. Y., Hittelman, J., Roth, J. A., Hong, W. K., Hittelman, W. N. (1994). Activation of p53 gene expression in premalignant lesions during head and neck tumorigenesis. Cancer research, 54(2), 321-326.

[18] Ambrosini, G., Adida, C., Altieri, D. C. (1997). A novel anti-apoptosis gene, survivin, expressed in cancer and lymphoma. Nature medicine, 3(8), 917-921.

[19] Thompson, C. B. (1995). Apoptosis in the pathogenesis and treatment of disease. Science, 267(5203), 1456-1462.

[20] Altieri, D. C. (2003). Survivin, versatile modulation of cell division and apoptosis in cancer. Oncogene, 22(53), 8581-8589.

[21] Missiaglia, E., Blaveri, E., Terris, B., Wang, Y. H., Costello, E., Neoptolemos, J. P., ... Lemoine, N. R. (2004). Analysis of gene expression in cancer cell lines identifies candidate markers for pancreatic tumorigenesis and metastasis. International journal of cancer, 112(1), 100-112.

[22] Shendure, J., Balasubramanian, S., Church, G. M., Gilbert, W., Rogers, J., Schloss, J. A., Waterston, R. H. (2017). DNA sequencing at 40: past, present and future. Nature, 550(7676), 345-353.

[23] Barba, M., Czosnek, H., Hadidi, A. (2014). Historical perspective, development and applications of next-generation sequencing in plant virology. Viruses, 6(1), 106–136. https://doi.org/10.3390/v6010106

[24] Mortazavi, A., Williams, B. A., McCue, K., Schaeffer, L., Wold, B. (2008). Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nature methods, 5(7), 621-628.

[25] Zhong, S., Joung, J. G., Zheng, Y., Chen, Y. R., Liu, B., Shao, Y., ... Giovannoni, J. J. (2011). High-throughput illumina strand-specific RNA sequencing library preparation. Cold spring harbor protocols, 2011(8), pdb-prot5652.

[26] Dobin, A., Gingeras, T. R. (2015). Mapping RNA-seq Reads with STAR. Current protocols in bioinformatics, 51, 11.14.1–11.14.19. https://doi.org/10.1002/0471250953.bi1114s51

[27] Conesa, A., Madrigal, P., Tarazona, S., Gomez-Cabrero, D., Cervera, A., McPherson, A., ... Mortazavi, A. (2016). A survey of best practices for RNA-seq data analysis. Genome biology, 17(1), 13.

[28] Dillies, M. A., Rau, A., Aubert, J., Hennequet-Antier, C., Jeanmougin, M., Servant, N., ... Guernec, G. (2013). A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Briefings in bioinformatics, 14(6), 671-683.

[29] Gutschner, T., Diederichs, S. (2012). The hallmarks of cancer: a long non-coding RNA point of view. RNA biology, 9(6), 703-719.

[30] Palazzo, A. F., Gregory, T. R. (2014). The case for junk DNA. PLoS Genet, 10(5), e1004351.

[31] Gutschner, T., Hämmerle, M., Eißmann, M., Hsu, J., Kim, Y., Hung, G., ... Zörnig, M. (2013). The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer research, 73(3), 1180-1189.

[32] Anastasiadou, E., Jacob, L. S., Slack, F. J. (2018). Non-coding RNA networks in cancer. Nature reviews. Cancer, 18(1), 5–18. https://doi.org/10.1038/nrc.2017.99

[33] Di Leva, G., Croce, C. M. (2013). miRNA profiling of cancer. Current opinion in genetics development, 23(1), 3-11.

[34] Rupaimoole, R., Slack, F. J. (2017). MicroRNA therapeutics: towards a new era for the management of cancer and other diseases. Nature reviews Drug discovery, 16(3), 203.

[35] Bartel, D. P. (2004). MicroRNAs: genomics, biogenesis, mechanism, and function. cell, 116(2), 281-297.

[36] Oberg, A. L., French, A. J., Sarver, A. L., Subramanian, S., Morlan, B. W., Riska, S. M., ... Smyrk, T. C. (2011). miRNA expression in colon polyps provides evidence for a multihit model of colon cancer. PloS one, 6(6), e20465.

[37] Calin, G. A., Dumitru, C. D., Shimizu, M., Bichi, R., Zupo, S., Noch, E., ... Rassenti, L. (2002). Frequent deletions and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia. Proceedings of the national academy of sciences, 99(24), 15524-15529.

[38] Dulak, A. M., Schumacher, S. E., Van Lieshout, J., Imamura, Y., Fox, C., Shim, B., ... Tabernero, J. (2012). Gastrointestinal adenocarcinomas of the esophagus, stomach, and colon exhibit distinct patterns of genome instability and oncogenesis. Cancer research, 72(17), 4383-4393.

[39] Tutar Y. (2012). Pseudogenes. Comparative and functional genomics, 2012, 424526. https://doi.org/10.1155/2012/424526

[40] Pei, B., Sisu, C., Frankish, A., Howald, C., Habegger, L., Mu, X. J., ... Reymond, A. (2012). The GENCODE pseudogene resource. Genome biology, 13(9), R51.

[41] Poliseno, L., Marranci, A., Pandolfi, P. P. (2015). Pseudogenes in Human Cancer. Frontiers in medicine, 2, 68. https://doi.org/10.3389/fmed.2015.00068

[42] Goodfellow, I., Bengio, Y., Courville, A., Bengio, Y. (2016). Deep learning (Vol. 1, p. 2). Cambridge: MIT press.

[43] Bishop, Christopher M. (2006). Pattern recognition and machine learning. New York: Springer.

[44] de Villiers, J., Barnard, E. (1993). Backpropagation neural nets with one and two hidden layers. IEEE Transactions on Neural Networks, 4(1), 136-141. doi:10.1109/72.182704

[45] Putri, O. (2018, December 21). Titanic Prediction with Artificial Neural Network in R. Retrieved November 04, 2020, from https://laptrinhx.com/titanic-prediction-with-artificial-neural-network-in-r-3087367370/.

[46] Li, M., Zhang, T., Chen, Y., Smola, A. J. (2014, August). Efficient mini-batch training for stochastic optimization. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 661-670).

[47] Anastassiou, G. A. (2011). Multivariate hyperbolic tangent neural network approximation. Computers Mathematics with Applications, 61(4), 809-821.

[48] Weisstein, Eric W. "Hyperbolic Tangent." From MathWorld–A Wolfram Web Resource. https://mathworld.wolfram.com/HyperbolicTangent.html

[49] Murugan, P. (2018). Implementation of deep convolutional neural network in multi-class categorical image classification. arXiv preprint arXiv:1801.01397.

[50] Rusiecki, A. (2019). Trimmed categorical cross-entropy for deep learning with label noise. Electronics Letters, 55(6), 319-320.

[51] Chollet, François. Keras. https://github.com/fchollet/keras, 2015.

[52] Glorot, X., Bengio, Y. (2010, March). Understanding the difficulty of training deep feedforward neural networks. In Proceedings of the thirteenth international conference on artificial intelligence and statistics (pp. 249-256).

[53] Larsen, J., Hansen, L. K. (1994). Generalization performance of regularized neural network models. Paper presented at the 42-51. doi:10.1109/NNSP.1994.366065

[54] Wang, S., Wang, X., Zhao, P., Wen, W., Kaeli, D., Chin, P., Lin, X. (2018, November). Defensive dropout for hardening deep neural networks under adversarial attacks. In Proceedings of the International Conference on Computer-Aided Design (pp. 1-8).

[55] Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., Salakhutdinov, R. (2014). Dropout: a simple way to prevent neural networks from overfitting. The journal of machine learning research, 15(1), 1929-1958.

[56] Buda, M., Maki, A., Mazurowski, M. A. (2018;2017;). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249-259. doi:10.1016/j.neunet.2018.07.011

[57] Zhou, Z., Liu, X. (2006). Training cost-sensitive neural networks with methods addressing the class imbalance problem. IEEE Transactions on Knowledge and Data Engineering, 18(1), 63-77. doi:10.1109/tkde.2006.17

[58] Johnson, J. M., Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. Journal of Big Data, 6(1), 1-54. doi:10.1186/s40537-019-0192-5

[59] Shrikumar, A., Greenside, P., Kundaje, A. (2017). Learning important features through propagating activation differences. arXiv preprint arXiv:1704.02685.

[60] Pleasance, E., Titmuss, E., Williamson, L., Kwan, H., Culibrk, L., Zhao, E. Y., ... Shen, Y. (2020). Pan-cancer analysis of advanced patient tumors reveals interactions between therapy and genomic landscapes. Nature Cancer, 1(4), 452-468.

[61] Grewal, J. K., Tessier-Cloutier, B., Jones, M., Gakkhar, S., Ma, Y., Moore, R., ... Lim, H. (2019). Application of a Neural Network Whole Transcriptome–Based Pan-Cancer Method for Diagnosis of Primary and Metastatic Cancers. JAMA network open, 2(4), e192597-e192597.

[62] Robinson, D. R., Wu, Y. M., Lonigro, R. J., Vats, P., Cobain, E., Everett, J., ... Schuetze, S. (2017). Integrative clinical genomics of metastatic cancer. Nature, 548(7667), 297-303.

[63] McKinney, W., others. (2010). Data structures for statistical computing in python. In Proceedings of the 9th Python in Science Conference (Vol. 445, pp. 51–56).

[64] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... Vanderplas, J. (2011). Scikit-learn: Machine learning in Python. The Journal of Machine Learning Research, 12, 2825-2830.

[65] Seo, Y., Shin, K. S. (2019). Hierarchical convolutional neural networks for fashion image classification. Expert Systems with Applications, 116, 328-339.

[66] Cancer Genome Atlas Research Network. (2017). Integrated genomic characterization of oesophageal carcinoma. Nature, 541(7636), 169-175.

[67] Cancer Genome Atlas Research Network. (2014). Comprehensive molecular characterization of gastric adenocarcinoma. Nature, 513(7517), 202-209.

[68] Zhang W. (2014). TCGA divides gastric cancer into four molecular subtypes: implications for individualized therapeutics. Chinese journal of cancer, 33(10), 469–470. https://doi.org/10.5732/cjc.014.10117

[69] Zeng, D., Li, M., Zhou, R., Zhang, J., Sun, H., Shi, M., ... Liao, W. (2019). Tumor microenvironment characterization in gastric cancer identifies prognostic and immunotherapeutically relevant gene signatures. Cancer immunology research, 7(5), 737-750.

[70] Buda, M., Maki, A., Mazurowski, M. A. (2018). A systematic study of the class imbalance problem in convolutional neural networks. Neural Networks, 106, 249-259.

[71] Guo, X., Yin, Y., Dong, C., Yang, G., Zhou, G. (2008, October). On the class imbalance problem. In 2008 Fourth international conference on natural computation (Vol. 4, pp. 192-201). IEEE.

[72] Japkowicz, N. (2000, June). The class imbalance problem: Significance and strategies. In Proc. of the Int'l Conf. on Artificial Intelligence (Vol. 56).

[73] Dudoit, S., Fridlyand, J., Speed, T. P. (2002). Comparison of discrimination methods for the classification of tumors using gene expression data. Journal of the American statistical association, 97(457), 77-87.

[74] Meyer, A. N., Payne, V. L., Meeks, D. W., Rao, R., Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: a vignette study. JAMA internal medicine, 173(21), 1952-1958.

[75] Anderson, G. G., Weiss, L. M. (2010). Determining tissue of origin for metastatic cancers: meta-analysis and literature review of immunohistochemistry performance. Applied immunohistochemistry molecular morphology : AIMM, 18(1), 3–8. https://doi.org/10.1097/PAI.0b013e3181a75e6d

[76] Sherman, B. T., Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols, 4(1), 44.

[77] Huang, D. W., Sherman, B. T., Lempicki, R. A. (2009). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research, 37(1), 1-13.

[78] Lewinski, A. (1988). Neuropeptides and Thyroid Function and Growth II. Intrathyroidal Peptidergic Nerves and Neuropeptides Located in Parafollicular (C) Cells. In Progress in Neuropeptide Research (pp. 65-72). Birkhäuser, Basel.

[79] Kanehisa, M., Goto, S. (2000). KEGG: kyoto encyclopedia of genes and genomes. Nucleic acids research, 28(1), 27-30.

[80] Kanehisa, M. (2019). Toward understanding the origin and evolution of cellular organisms. Protein Science, 28(11), 1947-1951.

[81] Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M., Tanabe, M. (2020). KEGG: integrating viruses and cellular organisms. Nucleic Acids Research.

[82] Patthy, L. (1991). Homology of the precursor of pulmonary surfactant-associated protein SP-B with prosaposin and sulfated glycoprotein 1. Journal of Biological Chemistry, 266(10), 6035-6037.

[83] Taylor-Papadimitriou, J., Stampfer, M., Bartek, J., Lewis, A., Boshell, M., Lane, E. B., Leigh, I. M. (1989). Keratin expression in human mammary epithelial cells cultured from normal and malignant tissue: relation to in vivo phenotypes and influence of medium. Journal of cell science, 94 (Pt 3), 403–413.

[84] Nanashima, N., Horie, K., Yamada, T., Shimizu, T., Tsuchida, S. (2017). Hair keratin KRT81 is expressed in normal and breast cancer cells and contributes to their invasiveness. Oncology reports, 37(5), 2964-2970.

[85] Lewis, M. T., Ross, S., Strickland, P. A., Snyder, C. J., Daniel, C. W. (1999). Regulated expression patterns of IRX-2, an Iroquois-class homeobox gene, in the human breast. Cell and tissue research, 296(3), 549-554.

[86] Chen, H., Sukumar, S. (2003). Role of homeobox genes in normal mammary gland development and breast tumorigenesis. Journal of mammary gland biology and neoplasia, 8(2), 159-175.

[87] Baffa, R., Fassan, M., Volinia, S., O'Hara, B., Liu, C. G., Palazzo, J. P., ... Rosenberg, A. (2009). MicroRNA expression profiling of human metastatic cancers identifies cancer gene targets. The Journal of Pathology: A Journal of the Pathological Society of Great Britain and Ireland, 219(2), 214-221.

[88] Pencheva, N., Tavazoie, S. F. (2013). Control of metastatic progression by microRNA regulatory networks. Nature cell biology, 15(6), 546-554.

[89] Bastide, A., David, A. (2018). The ribosome,(slow) beating heart of cancer (stem) cell. Oncogenesis, 7(4), 1-13.

[90] Artero-Castro, A., Kondoh, H., Fernandez-Marcos, P. J., Serrano, M., y Cajal, S. R., Lleonart, M. E. (2009). Rplp1 bypasses replicative senescence and contributes to transformation. Experimental cell research, 315(8), 1372-1383.

[91] Kim, J. H., You, K. R., Kim, I. H., Cho, B. H., Kim, C. Y., Kim, D. G. (2004). Overexpression of the ribosomal protein L36a gene is associated with cellular proliferation in hepatocellular carcinoma. Hepatology, 39(1), 129-138.

[92] Yang, S., Cui, J., Yang, Y., Liu, Z., Yan, H., Tang, C., ... Wang, W. (2016). Overexpressed RPL34 promotes malignant proliferation of non-small cell lung cancer cells. Gene, 576(1), 421-428.

[93] Zhou, H., Wang, Y., Lv, Q., Zhang, J., Wang, Q., Gao, F., ... Li, L. (2016). Overexpression of ribosomal RNA in the development of human cervical cancer is associated with rDNA promoter hypomethylation. PLoS One, 11(10), e0163340.

[94] Uemura, M., Zheng, Q., Koh, C. M., Nelson, W. G., Yegnasubramanian, S., De Marzo, A. M. (2012). Overexpression of ribosomal RNA in prostate cancer is common but not linked to rDNA promoter hypomethylation. Oncogene, 31(10), 1254-1263.

[95] Tsoi, H., Lam, K. C., Dong, Y., Zhang, X., Lee, C. K., Zhang, J., ... Fang, J. (2017). Pre-45s rRNA promotes colon cancer and is associated with poor survival of CRC patients. Oncogene, 36(44), 6109-6118.

[96] Ebright, R. Y., Lee, S., Wittner, B. S., Niederhoffer, K. L., Nicholson, B. T., Bardia, A., ... Mai, A. (2020). Deregulation of ribosomal protein expression and translation promotes breast cancer metastasis. Science, 367(6485), 1468-1473.

[97] Penzo, M., Montanaro, L., Treré, D., Derenzini, M. (2019). The Ribosome Biogenesis-Cancer Connection. Cells, 8(1), 55. https://doi.org/10.3390/cells8010055

[98] Catez, F., Dalla Venezia, N., Marcel, V., Zorbas, C., Lafontaine, D. L., Diaz, J. J. (2019). Ribosome biogenesis: An emerging druggable pathway for cancer therapeutics. Biochemical pharmacology, 159, 74-81.

[99] Kim, S. W., Roh, J., Lee, H. S., Ryu, M. H., Park, Y. S., Park, C. S. (2020). Expression of the immune checkpoint molecule V-set immunoglobulin domain-containing 4 is associated with poor prognosis in patients with advanced gastric cancer. Gastric Cancer, 1-14.

[100] Li, Y., Guo, M., Fu, Z., Wang, P., Zhang, Y., Gao, Y., Yue, M., Ning, S., Li, D. (2017). Immunoglobulin superfamily genes are novel prognostic biomarkers for breast cancer. Oncotarget, 8(2), 2444–2456. https://doi.org/10.18632/oncotarget.13683

[101] Kim, H. J., Maiti, P., Barrientos, A. (2017). Mitochondrial ribosomes in cancer. Seminars in cancer biology, 47, 67–81. https://doi.org/10.1016/j.semcancer.2017.04.004

[102] Lyng, H., Brøvig, R. S., Svendsrud, D. H., Holm, R., Kaalhus, O., Knutstad, K., ... Stokke, T. (2006). Gene expressions and copy numbers associated with metastatic phenotypes of uterine cervical cancer. BMC genomics, 7(1), 268.

[103] Caballero, O. L., Chen, Y. T. (2012). Cancer/testis antigens: potential targets for immunotherapy. In Innate Immune Regulation and Cancer Immunotherapy (pp. 347-369). Springer, New York, NY.

[104] Lahtz, C., Pfeifer, G. P. (2011). Epigenetic changes of DNA repair genes in cancer. Journal of molecular cell biology, 3(1), 51–58. https://doi.org/10.1093/jmcb/mjq053

[105] Kappil, M. A., Liao, Y., Terry, M. B., Santella, R. M. (2016). DNA Repair Gene Expression Levels as Indicators of Breast Cancer in the Breast Cancer Family Registry. Anticancer research, 36(8), 4039–4044.

[106] Sample, K. M. (2020). DNA repair gene expression is associated with differential prognosis between HPV16 and HPV18 positive cervical cancer patients following radiation therapy. Scientific reports, 10(1), 1-9.

[107] Gjerstorff, M. F., Terp, M. G., Hansen, M. B., Ditzel, H. J. (2016). The role of GAGE cancer/testis antigen in metastasis: the jury is still out. BMC cancer, 16, 7. https://doi.org/10.1186/s12885-015-1998-y

[108] Gjerstorff, M. F., Ditzel, H. J. (2008). An overview of the GAGE cancer/testis antigen family with the inclusion of newly identified members. Tissue antigens, 71(3), 187–192. https://doi.org/10.1111/j.1399-0039.2007.00997.x

[109] De Backer, O., Arden, K. C., Boretti, M., Vantomme, V., De Smet, C., Czekay, S., ... Van den Eynde, B. (1999). Characterization of the GAGE genes that are expressed in various human cancers and in normal testis. Cancer research, 59(13), 3157-3165.

[110] Chao, N. X., Li, L. Z., Luo, G. R., Zhong, W. G., Huang, R. S., Fan, R., Zhao, F. L. (2018). Cancer-testis antigen GAGE-1 expression and serum immunoreactivity in hepatocellular carcinoma. Nigerian journal of clinical practice, 21(10), 1361-1367.

[111] Shi, W. Y., Liu, K. D., Xu, S. G., Zhang, J. T., Yu, L. L., Xu, K. Q., Zhang, T. F. (2014). Gene expression analysis of lung cancer. Eur Rev Med Pharmacol Sci, 18(2), 217-28.

[112] Venugopal, N., Yeh, J., Kodeboyina, S. K., Lee, T. J., Sharma, S., Patel, N., Sharma, A. (2019). Differences in the early stage gene expression profiles of lung adenocarcinoma and lung squamous cell carcinoma. Oncology Letters, 18(6), 6572-6582.

[113] Parsons, C., Tayoun, A. M., Benado, B. D., Ragusa, G., Dorvil, R. F., Rourke, E. A., ... Habibian, M. (2018). The role of long noncoding RNAs in cancer metastasis. J Cancer Metastasis Treat, 4(4), 19.

[114] Adams, B. D., Parsons, C., Walker, L., Zhang, W. C., Slack, F. J. (2017). Targeting noncoding RNAs in disease. The Journal of clinical investigation, 127(3), 761-771.

[115] Adams, B. D., Slack, F. J. (2015). MicroRNA Signatures as Biomarkers in Cancer. eLS, 1-20.

[116] Huang T, Alvarez A, Hu B, Cheng SY. Noncoding RNAs in cancer and cancer stem cells. Chin J Cancer. 2013;32(11):582-593. doi:10.5732/cjc.013.10170

[117] Li, Y., Zheng, Q., Bao, C., Li, S., Guo, W., Zhao, J., ... Huang, S. (2015). Circular RNA is enriched and stable in exosomes: a promising biomarker for cancer diagnosis. Cell research, 25(8), 981-984.

[118] Hansen, T. B., Kjems, J., Damgaard, C. K. (2013). Circular RNA and miR-7 in cancer. Cancer research, 73(18), 5609-5612.

[119] Lu, J., Getz, G., Miska, E. A., Alvarez-Saavedra, E., Lamb, J., Peck, D., ... Downing, J. R. (2005). MicroRNA expression profiles classify human cancers. nature, 435(7043), 834-838.

[120] Sandhu, S. K., Croce, C. M., Garzon, R. (2011). Micro-RNA expression and function in lymphomas. Advances in hematology, 2011.

[121] Navarro, A., Gaya, A., Martinez, A., Urbano-Ispizua, A., Pons, A., Balagué, O., ... Montserrat, E. (2008). MicroRNA expression profiling in classic Hodgkin lymphoma. Blood, The Journal of the American Society of Hematology, 111(5), 2825-2832.

[122] Zhang, J., Jima, D. D., Jacobs, C., Fischer, R., Gottwein, E., Huang, G., ... Weinberg, J. B. (2009). Patterns of microRNA expression characterize stages of human B-cell differentiation. Blood, 113(19), 4586-4594.

[123] Li, C., Kim, S. W., Rai, D., Bolla, A. R., Adhvaryu, S., Kinney, M. C., ... Aguiar, R. C. (2009). Copy number abnormalities, MYC activity, and the genetic fingerprint of normal B cells mechanistically define the microRNA profile of diffuse large B-cell lymphoma. Blood, The Journal of the American Society of Hematology, 113(26), 6681-6690.

[124] Baranwal S, Alahari SK. miRNA control of tumor cell invasion and metastasis. Int J Cancer. 2010;126(6):1283-1290. doi:10.1002/ijc.25014

[125] Singh, R., Saini, N. (2012). Downregulation of BCL2 by miRNAs augments drug-induced apoptosis–a combined computational and experimental approach. Journal of cell science, 125(6), 1568-1578.

[126] Pekarsky, Y., Balatti, V., Croce, C. M. (2018). BCL2 and miR-15/16: from gene discovery to treatment. Cell death and differentiation, 25(1), 21–26. https://doi.org/10.1038/cdd.2017.159

[127] Roehle, A., Hoefig, K. P., Repsilber, D., Thorns, C., Ziepert, M., Wesche, K. O., ... Matolcsy, A. (2008). MicroRNA signatures characterize diffuse large B-cell lymphomas and follicular lymphomas. British journal of haematology, 142(5), 732-744.

[128] Li, J., Zou, J., Wan, X., Sun, C., Peng, F., Chu, Z., Hu, Y. (2020). The Role of Noncoding RNAs in B-Cell Lymphoma. Frontiers in oncology, 10, 577890. https://doi.org/10.3389/fonc.2020.577890

[129] Balatti V, Pekarky Y, Croce CM. Role of microRNA in chronic lymphocytic leukemia onset and progression. J Hematol Oncol. 2015;8:12. Published 2015 Feb 20. doi:10.1186/s13045-015-0112-x

[130] Mollashahi B, Aghamaleki FS, Movafagh A. The Roles of miRNAs in Medulloblastoma: A Systematic Review. J Cancer Prev. 2019;24(2):79-90. doi:10.15430/JCP.2019.24.2.79

[131] Cho, W. C. (2010). MicroRNAs: potential biomarkers for cancer diagnosis, prognosis and targets for therapy. The international journal of biochemistry cell biology, 42(8), 1273-1281.

[132] Joshi, P., Katsushima, K., Zhou, R., Meoded, A., Stapleton, S., Jallo, G., ... Perera, R. J. (2019). The therapeutic and diagnostic potential of regulatory noncoding RNAs in medulloblastoma. Neuro-oncology advances, 1(1), vdz023.

[133] Mollashahi B, Aghamaleki FS, Movafagh A. The Roles of miRNAs in Medulloblastoma: A Systematic Review. J Cancer Prev. 2019;24(2):79-90. doi:10.15430/JCP.2019.24.2.79

[134] Joshi, P., Katsushima, K., Zhou, R., Meoded, A., Stapleton, S., Jallo, G., ... Perera, R. J. (2019). The therapeutic and diagnostic potential of regulatory noncoding RNAs in medulloblastoma. Neuro-oncology advances, 1(1), vdz023.

[135] Zheng, B., Xi, Z., Liu, R., Yin, W., Sui, Z., Ren, B., ... Liu, C. (2018). The function of microRNAs in B-cell development, lymphoma, and their potential in clinical practice. Frontiers in immunology, 9, 936.

[136] Poliseno, L., Marranci, A., Pandolfi, P. P. (2015). Pseudogenes in Human Cancer. Frontiers in medicine, 2, 68. https://doi.org/10.3389/fmed.2015.00068

[137] Gao, K. M., Chen, X. C., Zhang, J. X., Wang, Y., Yan, W., You, Y. P. (2015). A pseudogene-signature in glioma predicts survival. Journal of experimental clinical cancer research, 34(1), 23.

[138] Welch, J. D., Baran-Gale, J., Perou, C. M., Sethupathy, P., Prins, J. F. (2015). Pseudogenes transcribed in breast invasive carcinoma show subtype-specific expression and ceRNA potential. BMC genomics, 16(1), 113.

[139] Zhang, Y., Parmigiani, G., Johnson, W. E. (2020). ComBat-Seq: batch effect adjustment for RNA-Seq count data. bioRxiv.