

Transcriptomic Profile Based Cancer Disease Prediction and Patient Survival Time Differentiation

by

Samuel Ofoosu Mensah



*Thesis presented in partial fulfillment of the requirements
for the degree of Master of Science in Mathematics in the
Faculty of Science at Stellenbosch University*



Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.

Supervisor: Dr Gaston Kuzamunu Mazandu & Co-supervised: Dr Simukai Wanziru Utete

December 2018

Declaration

By submitting this thesis electronically, I declare that the entirety of the work contained therein is my own, original work, that I am the sole author thereof (save to the extent explicitly otherwise stated), that reproduction and publication thereof by Stellenbosch University will not infringe any third party rights and that I have not previously in its entirety or in part submitted it for obtaining any qualification.

Signature:
Samuel Ofosu Mensah

Date: December 2018

Copyright © 2018 Stellenbosch University
All rights reserved.

Abstract

Transcriptomic Profile Based Cancer Disease Prediction and Patient Survival Time Differentiation

Samuel Oforu Mensah

*Department of Mathematical Sciences,
University of Stellenbosch,
Private Bag X1, Matieland 7602, South Africa.*

Thesis: MSc. (Mathematics)

December 2018

Cancer disease is an abnormal growth of cells, which may be caused by mutations in genes which, as a result, alter the way cells function mainly in the way they grow and divide. Cancer cells are regulated by complex interactions mediated by a group of proteins and miRNAs which are expressed and repressed. With the help of transcriptomic technologies such as RNA–sequencing (RNA–seq), it is now possible to profile thousands of genes at once to create a global picture of the functions of cells. Here, the study employs a statistical approach, called Significance Analysis of Microarray (SAM), to identify genes that are differentially expressed in breast cancer patients. Genes with scores greater than a threshold are deemed potentially significant. Genes identified as significantly different are used for twofold reasons. First, the study uses these significantly identified genes to predict breast cancer using three machine learning algorithms. The machine learning algorithms used are random forests, artificial neural networks and support vector machines. Secondly, clinical details of patients and significantly identified genes are combined to build a survival model to predict the probability of survival and risk to the event in breast cancer patients. Using The Cancer Genome Atlas (TCGA) as the primary data for

the study, SAM reported 23 genes as significantly different. Further investigations revealed that these 23 significant genes are involved in tumour suppression, angiogenesis, cell growth factor, tumourigenesis, cell proliferation, tumour progression and tumour necrosis activities. In predicting breast cancer, 10 out of the 23 genes contribute significantly to the model. Finally, it was identified that log–logistic distribution best describes the survival time of breast cancer patients. Moreover, the survival model revealed that expression levels of six genes influence the survival probability of a breast cancer patient.

Keywords: Breast Cancer, RNA–sequencing, Differential Expression, Significance Analysis of Microarray, TCGA, Machine Learning, Survival Analysis.

Opsomming

Transcriptomic Profiel Gebaseer Kanker Siekte Voorspelling en Geduldige Oorlewing Tyd Onderskeid

(“Transcriptomic Profile Based Cancer Disease Prediction and Patient Survival Time Differentiation ”)

Samuel Ofosu Mensah

*Departement Wiskundige Wetenskappe,
Universiteit van Stellenbosch,
Privaatsak X1, Matieland 7602, Suid Afrika.*

Tesis: MSc. (Wiskunde)

Desember 2018

Kanker siekte is 'n abnormale groei van selle, wat veroorsaak kan word deur mutasies in gene, gevolglik, verander die manier waarop selle hoofsaaklik funksioneer in die manier waarop hulle groei en verdeel. Kanker selle word geregleer deur komplekse interaksies gemedieer deur 'n groep proteiene en miRNAs wat uitgedruk en onderdruk word. Met behulp van transcriptomiese tegnologie soos RNA-sequencing (RNA - seq), is dit nou moontlik om duisende gene gelyktydig te profileer om 'n globale prentjie van die funksies van selle te skep. Hier gebruik die studie 'n statistiese benadering, genoem Significance Analysis of Microarray (SAM), om betekenisvolle gene te identifiseer wat differensieel uitgedruk word in borskanker pasiënte. Genes met tellings groter as 'n drempel word beskou as potensieel betekenisvol. Vervolgens gebruik die studie hierdie beduidende geïdentifiseerde gene om borskanker te voorspel deur gebruik te maak van drie machine learning algoritmes, insluitend random forests, artificial neural networks en support vector machines. Laastens word kliniese besonderhede van pasiënte en beduidende geïdentifiseerde gene gekombineer om 'n

oorlewingsmodel te bou om die waarskynlikheid van oorlewing en risiko vir die gebeurtenis in pasiënte met borskanker te voorspel. Die risiko vir die geleentheid vir hierdie studie is die dood. Met behulp van The Cancer Genome Atlas (TCGA) as die primêre data vir die studie, het SAM 23 gene so beduidend anders aangedui. Verdere ondersoek het getoon dat hierdie 23 belangrike gene betrokke was by tumour suppression, angiogenesis, sel groeifaktor, tumorigenesis, sel proliferasie, tumor progressie en tumor necrosis aktiwiteite. By die voorspel van borskanker dra 10 uit die 23 gene aansienlik by tot die model. Ten slotte is geïdentifiseer dat log-logistieke verspreiding die oorlewingstyd van pasiënte met borskanker die beste beskryf. Daarbenewens het die oorlewingsmodel geopenbaar dat uitdrukkingsvlakke van ses gene die oorlewingswaarskynlikheid van 'n pasiënt met borskanker beïnvloed. Die oorlewingsmodel het verder getoon dat borskanker pasiënte waarskynlik groter risiko vir die gebeurtenis sal hê, maar na 3243.38 dae kan hul risiko vir die gebeurtenis geleidelik verminder.

Keywords: RNA-Seq, Borskanker, Differentiële Uitdrukking, Betekenisanalise van Microarray, Masjienleer, Oorlewingsontleding.

Acknowledgements

This project had been a successful one as a result of guidance, encouragement and support from many. I, first of all, acknowledge my success to the Almighty God for His unchanged grace and a sense of direction throughout the period of the research.

I also wish to acknowledge my supervisor, Dr Gaston Kuzamunu Mazandu for his support, guidance, and sacrifice made in making this project a success.

I would also like to express my heartfelt gratitude to my co-supervisor Dr Simukai W. Utete, for her kind gesture, guidance and sacrifices made in making this project a success.

Not forgetting my profound gratitude to the African Institute for Mathematical Sciences (AIMS) for giving me this opportunity to study at their research centre.

To all who gave me advice in diverse ways most especially to my parents, Mr and Mrs Emmanuel K. Mensah and my siblings for their love and encouragement towards this project, I say thank you. I also extend my sincere gratitude to all my noble friends.

Lastly, I would like to thank the members of the Lighthouse Chapel International especially to Bishop Dag Heward-Mills and Rev. Napoleon Essien for their prayers and guidance. To all I say, thank you and God richly bless you.

Dedications

I dedicate this work to my dear parents Mr and Mrs Emmanuel K. Mensah.

Contents

Declaration	i
Abstract	ii
Opsomming	iv
List of Figures	x
List of Tables	xi
1 Introduction	1
1.1 Background	1
1.2 Problem statement	3
1.3 Aims and objectives of the study	3
1.4 Contributions of the study	4
1.5 Outline of the study	4
2 Literature review	5
2.1 A brief review of RNA–sequencing	5
2.2 A brief review of significance analysis of microarray (SAM)	7
2.3 A brief review of classification algorithms in medical studies	8
2.4 A brief review of survival analysis	9
3 Different analysis method for breast cancer	11
3.1 RNA–seq data matrix	11
3.2 Describing RNA–seq data used	13
3.3 Extraction of essential genes	13
3.4 Exploring different classification algorithms	17

3.4.1	Explaining random forests	17
3.4.2	Explaining artificial neural network (ANN)	21
3.4.3	Support vector machines (SVM)	28
3.4.4	Median–supplement: a balancing data technique	30
3.5	Performance metrics for classification algorithms	32
3.6	Investigating survival analysis	34
3.6.1	Defining survival function	35
3.6.2	Exploring non–parametric survival models	35
3.6.3	Exploring specific probability distributions	37
3.6.4	Semi-parametric survival models	48
3.6.5	Statistical model selection	48
3.7	Software and packages used	50
4	Statistical analysis of breast cancer gene expression and clinical data	51
4.1	Identifying essential genes	51
4.2	Classifying patients based on essential genes	54
4.2.1	Comparing the performance of classification algorithms	55
4.2.2	Details of the best performing classification algorithm	57
4.3	Association between essential genes and patients survival	59
4.3.1	Kaplan–Meier survival analysis	60
4.3.2	Building parametric survival models	64
4.3.3	Analysing the Cox proportional hazard model	70
4.4	Discussion of results	72
5	Conclusion	78
	List of references	80

List of Figures

2.1	Number of articles published about transcriptomic technology	6
3.1	Pipeline for obtaining RNA-seq	12
3.2	An example of a decision tree	18
3.3	Mathematically modelling ANN using biological neuron	22
3.4	Design of an Artificial Neural Network	23
3.5	Effect of learning rates on loss function	26
3.6	A cost function with two minima	27
3.7	Seperating two groups using SVM	28
3.8	A Latin hypercube sampling	31
3.9	A confusion matrix	33
3.10	Plots for exponential distribution	41
3.11	Plots for Weibull distribution	42
3.12	Plots for log-logistic distribution	43
3.13	Plots for log-normal distribution	45
4.1	Q-Q plot from the SAM algorithm	52
4.2	Variable importance of the genes in random forest	58
4.3	A tree obtained from the random forest algorithm	59
4.4	Right censored survival plot for 50 patients	60
4.5	Kaplan-Meier survival curve for patients	61
4.6	Kaplan-Meier survival curves for two-risk group	63
4.7	Comparing Kaplan-Meier curve with probability distributions	65
4.8	Log-logistic distribution plots	69

List of Tables

4.1	Significant genes with their corresponding description and scores . . .	53
4.2	Molecular function and biological process for significant genes	54
4.3	Performance of the classifiers without supplement data	56
4.4	Performance of the classifiers using median supplement data	56
4.5	Performance of the classifiers using mean supplement data	57
4.6	Log-rank test statistic for genes	62
4.7	Probability distributions with their corresponding AIC values	65
4.8	Details of log-logistic model with the 23 significant genes	66
4.9	Variable Selection Process for log-logistic distribution	67
4.10	Details of the reduced log-logistic model	68
4.11	Details of Cox Proportional Hazard with the 23 significant genes . . .	70
4.12	Variable Selection Process for Cox Proportional Hazard	71
4.13	Details of the reduced Cox Proportional Hazard	71

Chapter 1

Introduction

1.1 Background

Cancer disease is an abnormal growth of cells, which may be caused by mutations in genes which, as a result, alter the way cells function mainly in the way they grow and divide ([National Cancer Institute, 2017](#)). Mankind has a history of experiencing different deadly diseases, but cancer has been considered as the most complex he has ever faced. According to the [World Health Organisation \(2017\)](#), almost 1 of 6 deaths is caused by cancer, making it the second leading causes of death in the world. [Tomczak *et al.* \(2015\)](#) stated that there are more than 200 forms of cancer discovered and each uniquely identified by a different molecular structure. In this study, we investigate breast cancer, which is the most common type of cancer in women ([Wang *et al.*, 2018](#)). Breast cancer is also the second cause of cancer death in women after lung cancer ([Siegel *et al.*, 2018](#)).

A conventional way of detecting breast cancer in women is by screening the breast organ using mammography. This approach of diagnosing breast cancer is however limited because of low sensitivity and specificity. For this reason, continuous research is being done to develop novel diagnostic and therapeutic strategies to improve breast cancer detection and treatment in women. Unfortunately, the molecular mechanisms involved in the formation and development of breast cancer continues to be ill-defined. Hence, it is essential to find novel genes that contribute to the formation and development of breast cancer ([Wang *et al.*, 2018](#)).

Nowadays transcriptomic technologies provide an opportunity for detecting genes that may influence breast cancer formation and development. These technologies are used to profile the genome of a cell which eventually measures the transcription level of the genes. A more accurate transcriptomic technology is the RNA–sequencing (RNA–seq). It is sensitive and provides a global picture of the functions of cells (Wang *et al.*, 2009).

Typically, RNA–seq from different cells can be combined to generate large matrices of breast cancer gene expression data. A comparative study may then be used to elucidate the differences that exist between the gene expression of breast cancer patients and normal samples. Genes that are significantly different in expression levels are known as differentially expressed genes. Usually, these comparative studies employ statistical tools, such as edgeR, DESeq2, SAMseq and many others, to identify differentially expressed genes (Li and Tibshirani, 2013). Using a statistical approach to detect genes that are significantly different have several advantages. They include improvement in classification related tasks, decreasing clinical cost and increasing biological knowledge of a disease (Jiangeng *et al.*, 2007).

There may exist elusive patterns in the breast cancer gene expression data. Moreover, gene expression data may have high dimensions and may be noisy. Using such data to make the predictions can be challenging. Machine learning methods are used to extract relevant features and train models to identify hidden patterns that exist in gene expression data to make accurate predictions (Danaee *et al.*, 2017).

An important information for clinicians is the survival details of a breast cancer patient. Traditionally, medical practitioners use cancer survival rates to predict the survival time of patients. In particular, the most common survival rate used by medical practitioners is the 5–year survival rate. However, this technique may not be a useful measure for developing prognostic tools and therapeutic interventions for breast cancer (Li *et al.*, 2017). Survival analysis techniques are therefore used to build models to predict patients' survival.

1.2 Problem statement

The biological mechanism involved in breast cancer is still unclear. Moreover, clinically examining a patient's genome to identify differentially expressed genes can be costly. To this end, recent advances in transcriptomic technology have been developed to provide essential tools to maximise meaningful insight into the genome of a cell (Wang *et al.*, 2009). However, data obtained from transcriptomic technologies can be high dimensional. This is because, the human genome contains about 20500 genes (Clamp *et al.*, 2007; Ezkurdia *et al.*, 2014). Even though a large amount of data is generated, methods are needed to identify significant genes that may contribute to breast cancer.

Earlier studies applied clustering techniques on the data generated to determine genes that are similar (van 't Veer *et al.*, 2002). Similar genes are then used to develop strategies for the treatment of breast cancer. For example, van 't Veer *et al.* (2002) used a hierarchical clustering algorithm to cluster 98 tumours based on their similarity. Other studies performed further investigations on genes that have already been identified as significant genes. For example, Mensah *et al.* (2017) clustered 969 breast cancer samples using genes that have already been identified as significant to breast cancer. These approaches may be useful but provide little information and may be limited in identifying novel genes (Tusher *et al.*, 2001). This study is conducted in the perspective of the above.

1.3 Aims and objectives of the study

The aims and objectives of this study are:

- i. To conduct a comparative study on patients in order to identify genes that are differentially expressed between breast cancer and normal samples.
- ii. To build a model to predict breast cancer based on a patient's transcription profile.
- iii. To build a model to predict the survival probability of a breast cancer patient based on expression levels of significantly identified genes.

1.4 Contributions of the study

A statistical technique called significance analysis of microarray (SAM) is used to identify differentially expressed genes. Earlier research work apply SAM only with a subset of the genome, as they have *a priori* knowledge about the significance of genes. Here we consider all genes of a genome to identify differentially expressed genes hence we have no such knowledge of which gene is significant. By so doing, we identified novel genes that contribute to the formation of breast cancer. We then use identified genes as features for supervised learning techniques to predict breast cancer. Furthermore, the identified genes are used to build a model to predict the survival probability of breast cancer patients.

1.5 Outline of the study

The rest of this thesis is organized as follows. Chapter 2 reviews the techniques used for the study. The techniques used for the study include significance analysis of microarray (SAM), three machine learning algorithms and survival analysis techniques. This Chapter also provides vivid reasons for employing these techniques. Chapter 3 presents a detailed description of the techniques used for the study. Chapter 4 also presents and discusses the results obtained in the study. Finally, Chapter 5 concludes and presents potential future work.

Chapter 2

Literature review

This chapter reviews important techniques used to analyse a patient's gene expression and associated clinical data.

2.1 A brief review of RNA–sequencing

Recent advancement in transcriptomic technologies has made it possible to study the entire genome of an organism at a large–scale (Poulin and Nielsen, 2009). These technologies have mainly been categorised into hybridization and sequencing methods (Wang *et al.*, 2009). More generally, hybridization methods quantify an organisms' transcriptome using a measure of fluorescence. A popular example of the hybridization method is microarray. On the other hand, sequencing methods use high–throughput approaches to deduce the transcriptome of an organism. A recent example of sequencing methods is the RNA–sequencing (RNA–seq). This review focuses on RNA–seq and microarray as they are the main contemporary technologies used in this area of research (Lowe *et al.*, 2017).

RNA–seq, however, has several advantages over microarray experiments. For example, Wang *et al.* (2009) mentioned that hybridization techniques rely on existing knowledge of a biological sample, produces high background noise and have a limited range of detection. Unlike microarray which makes use of hybridization, RNA–seq can produce transcripts with or without a reference genome of a biological sample. For this reason, it is the ideal technique to be used if the objective of a transcription profiling experiment is to identify novel genes. Also, it is relatively capable of mapping DNA sequences to unique re-

gions of a genome with low background noise. Finally, it has no upper limit for quantification of DNA sequences identified, hence making it have a wider range of detecting expressions. Thus, RNA-seq is capable of identifying more differentially expressed genes with higher fold change (Zhao *et al.*, 2014). Also, Costa-Silva *et al.* (2017) stated that RNA-seq is highly reproducible with a little number of technical replicates. However, microarrays are likely to face a problem called cross-hybridization, which involves an overlap of fluorescence dyes reducing the signal intensities of a transcript (Draghici *et al.*, 2006).

RNA-seq also permits quantitative profiling and is, therefore, increasing our knowledge of the transcriptome. In general, RNA-seq has provided increased detection sensitivity and has open new avenues of research in transcriptome analyses, such as the study of gene fusions, allele-specific expression and novel alternative transcripts (Su *et al.*, 2014). In this regard, RNA-seq is continuously becoming the method of choice for transcriptional profiling experiments (Corney and Basturea, 2013). To illustrate further, Figure 2.1 shows the history of how these two technologies have been used for transcriptomic profiling.

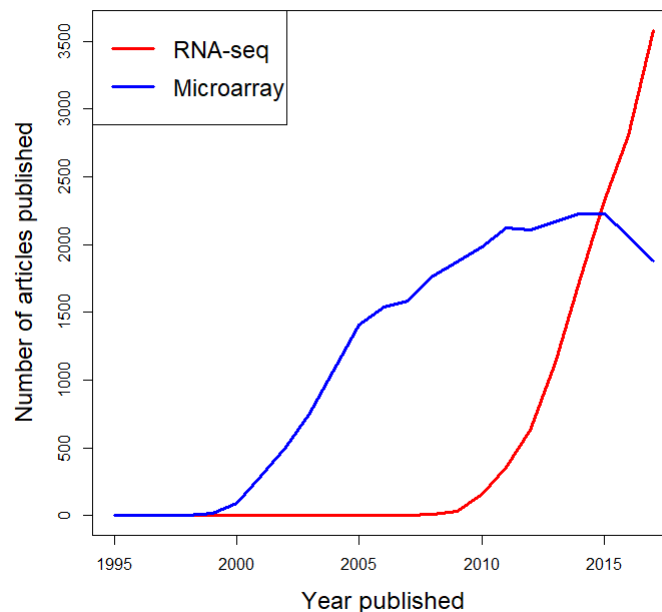


Figure 2.1: A history of the number of articles published about a transcriptomic technology. Microarray used to be the preferred transcriptomic technology but it use declines after RNA-seq was introduced.

In general, the RNA-seq technology is very useful for differential expression analysis involving some specific conditions (Costa-Silva *et al.*, 2017). Zhang *et al.* (2014) also noted that the cost involved in RNA-seq experiments is gradually decreasing due to its continuous interest as an ideal technology for transcriptomic profiling. In summary, RNA-seq is a sequencing approach that can quantitatively profile the entire transcriptome of a biological sample in a high-throughput manner (Wang *et al.*, 2009).

2.2 A brief review of significance analysis of microarray (SAM)

Many statistical methods have been developed to analyse RNA-seq data, due to its growing popularity in profiling experiments (Li and Tibshirani, 2013; Li and Li, 2018; Seyednasrollah *et al.*, 2013). Li and Li (2018) mentioned that more than two thousand of these statistical methods have been designed in the past decade to help visualise, process, analyse and interpret RNA-seq data. These statistical methods specifically help to identify genes with changes in the level of expression between comparison groups (Poulin and Nielsen, 2009).

Li and Tibshirani (2013) reported that the statistical methods may either be parametric or non-parametric. Parametric methods assume that gene expression follows underlying distributions, such as normal, negative binomial, or Poisson distribution. Most commonly used parametric methods includes edgeR (Robinson *et al.*, 2009), DESeq (Anders and Huber, 2010; Love *et al.*, 2014), Poisson-Seq (Li *et al.*, 2012), and baySeq (Hardcastle and Kelly, 2010). Even though the sample size of the experiment may be small, parametric models can be powerful and efficient only when their underlying assumptions hold. However, results from parametric models can be unreliable for real data, because it is uncertain to be correctly described by the assumed distribution (Li and Tibshirani, 2013). Non-parametric methods, however, do not assume that gene expression follows an underlying distribution, but are capable of producing reliable results. An example of the non-parametric method used to identify differential genes is the SAM-seq (Li and Tibshirani, 2013).

Here, we use SAM-seq to identify differentially expressed genes. This method

of differential gene expression is used because, RNA-seq data is in the form of counts and can be very skewed due to outliers. Parametric approaches, unfortunately, are sensitive to outliers making them not the best techniques to answer questions related to differential gene expressions. In addition, SAM-seq can handle several output types, including one-class, survival, quantitative and multiple-class. Also, SAM-seq incorporates a resampling technique in the algorithm to get rid of sequencing depth problems which emanate by reason of different ways of performing experiments (Li and Tibshirani, 2013). SAM-seq as a non-parametric approach may have several advantages over parametric methods, but there is no principled guide for best practices (Seyednasrollah *et al.*, 2013).

Numerous studies have been carried out to compare the performance of these statistical methods. Li and Tibshirani (2013) simulated three types of data (first data contained outliers, second data did not contain any outlier, and the third data had a small sample size) to evaluate the performance of edgeR, DESeq, PoissonSeq, and SAM-seq. They further compared these statistical methods using three different real data, namely; Marioni data, t'Hoen data, and Witten data. For all, they noted significant genes identified by parametric methods were extremely influenced by outliers whereas SAM-seq was robust with outliers, simple to use and consistent with identified significant genes. Similarly, Seyednasrollah *et al.* (2013) compared 8 different statistical methods to formulate principled guidelines for best practices. Zhang *et al.* (2014) also carried out a comparative study of techniques used for RNA-seq differential expression analysis. Likewise, Sonesson and Delorenzi (2013) conducted a comparative study on 11 statistical methods used for RNA-seq differential analysis. They concluded that, for data with large sample sizes, SAM-seq performed well under several conditions.

2.3 A brief review of classification algorithms in medical studies

Lately, machine learning algorithms have become essential tools in medical diagnosis primarily because they are effective in helping clinicians to classify and

recognise diseases (Polat and Güneş, 2007). As an illustration, a machine learning algorithm that correctly classifies cancer patients to their corresponding classes can effectively guide medical experts in making therapeutic decisions (Vanneschi *et al.*, 2011). Recent achievements in medical data have led to high-dimensionality in feature space problems, linearly inseparable and missing data issues. Hence, the need for machine learning techniques to fully extract information from these data (Khondoker *et al.*, 2016).

Several studies compare different machine learning techniques to assess their performance to make accurate decisions (Kourou *et al.*, 2015). Such comparisons are done mainly because of the “no free lunch theorem” which states that there is no single learning algorithm that universally performs best across all domains (Wolpert and Macready, 1995). In addition, different machine learning may be employed based on the type of data. For example, support vector machines may be used in cases where data is linearly inseparable (Friedman *et al.*, 2001).

For this reason, different machine learning techniques may be explored (Douglas *et al.*, 2011). Studies like Douglas *et al.* (2011) compared the accuracy of six different machine learning algorithms using neuroimaging data. Similarly, Khondoker *et al.* (2016) compared five widely used machine learning techniques on simulated data to investigate the performance of the algorithms under different circumstances. In addition, Yue *et al.* (2018) employed four machine learning techniques to correctly diagnose breast cancer patients using the “Wisconsin breast cancer database” as their primary data. Nevertheless, various studies reported different best performing algorithms which justifies the “no free lunch theorem”. Here, we used random forests, artificial neural network (ANN), and support vector machines (SVM) because they are able to handle data with high dimensional feature space problems and data that are not linearly separable.

2.4 A brief review of survival analysis

One major task of clinicians is to make meaningful interpretation and prediction of patients molecular information. Hence, there is a high demand to determine the outcome of patients based on their molecular profiles. Identifying the relationship that exists between clinical results and molecular information has led

to the adaptation of survival analysis to help determine a patient's outcome. In the past decade, molecular information has assisted in the identification of prognostic factors and therapeutic targets, hence the need to include them in survival analysis techniques (Chen *et al.*, 2014).

Mostly, clinical information consists of the patient's survival details which may include censoring. The use of statistical approaches, such as linear and logistic regression cannot account for the censoring in a clinical information (McGready, 2009). Hence, several medical research has employed survival analysis to analyse a patient's outcome. For example, Abadi *et al.* (2014) used survival analysis to investigate how different treatment of breast cancer influence their survival time. Specifically, they used the Cox proportional hazard model to study 15830 women diagnosed with breast cancer. Their results indicated that radiotherapy and chemotherapy increased the hazard of patients at the first and second stage of breast cancer.

Also, Carey *et al.* (2006) investigated the prevalence of breast cancer subtypes within racial subsets to determine their association with breast cancer survival. They discovered that young African American patients had a higher prevalence of breast tumours compared to older African Americans and non-American patients. They, therefore, concluded that a higher prevalence of basal-like breast tumours and lower prevalence of luminal A tumours may influence the poor prognosis of young African American women with breast cancer. Miecznikowski *et al.* (2010) also developed a model to determine the survival of breast cancer patients using their gene expression data. In particular, they use Cox proportional hazard to identify tumour size and oestrogen status as the main influencer of breast cancer.

In short, there have been several applications and developments in survival analysis. For instance, Liang *et al.* (2016) developed a novel semi-supervised learning method founded on Cox proportional hazard and accelerated failure time model to predict risk involved in treatment and survival time of patients. Chai *et al.* (2017) identified that the model is easily influenced by noise. Hence, they improved it by incorporating a self-paced learning algorithm to fully utilize censored data. Here, our major interest is in using significant genes identified by significance analysis of microarray (SAM) to build a model to predict the survival probability of breast cancer patients.

Chapter 3

Different analysis method for breast cancer gene expression and clinical data

This chapter gives details of techniques and approaches used for this research. First of all, it gives details of the type of data used for the study and how it is generated. This is then followed by a technique used to find genes that are differentially expressed. Next, we explore three machine learning methods to find the best classification algorithm for the study. Finally, survival analysis is employed to build a model and predict the survival time for patients expressing particular genes.

3.1 RNA-seq data matrix

Figure 3.1 shows the process for obtaining RNA-seq data. To begin with, RNA-seq uses Next Generation Sequencing (NGS) to profile the transcriptome of an organism. The first step of RNA-seq is to randomly fragment messenger RNA (mRNA) into short pieces. This is followed by a process called reverse transcription which involves converting the fragmented mRNA to complementary DNA (cDNA) using a random primer. With the help of Polymerase chain reaction (PCR) amplification, cDNA is used to generate millions of short sequence reads. Finally, mapping algorithms are used unambiguously to identify the region where these short reads belong ([Wang *et al.*, 2009](#)). A popular algo-

rithm used for mapping short sequence reads to the region of interest is the RNA–sequencing Expectation Maximization (RSEM). Specifically, RSEM uses a Bayesian network approach to estimate the abundant level of a gene (Li and Dewey, 2011; Li *et al.*, 2009). This protocol has been implemented in sequencing machines. For instance, the sequencing machine used to generate data for this study is called Illumina HiSeq. Note that the process described above is only for one sample. Fortunately, sequencing machines allow multiple samples to be run in parallel. Results from running multiple samples can, therefore, be summarised to form a matrix.

Suppose there are n samples, each having p number of genes, then the resulting matrix N with n rows and p columns ($n \times p$ dimension). Element N_{ij} of matrix N is the number of reads mapped to gene j , for $1 \leq j \leq p$ in sample i , for $1 \leq i \leq n$.

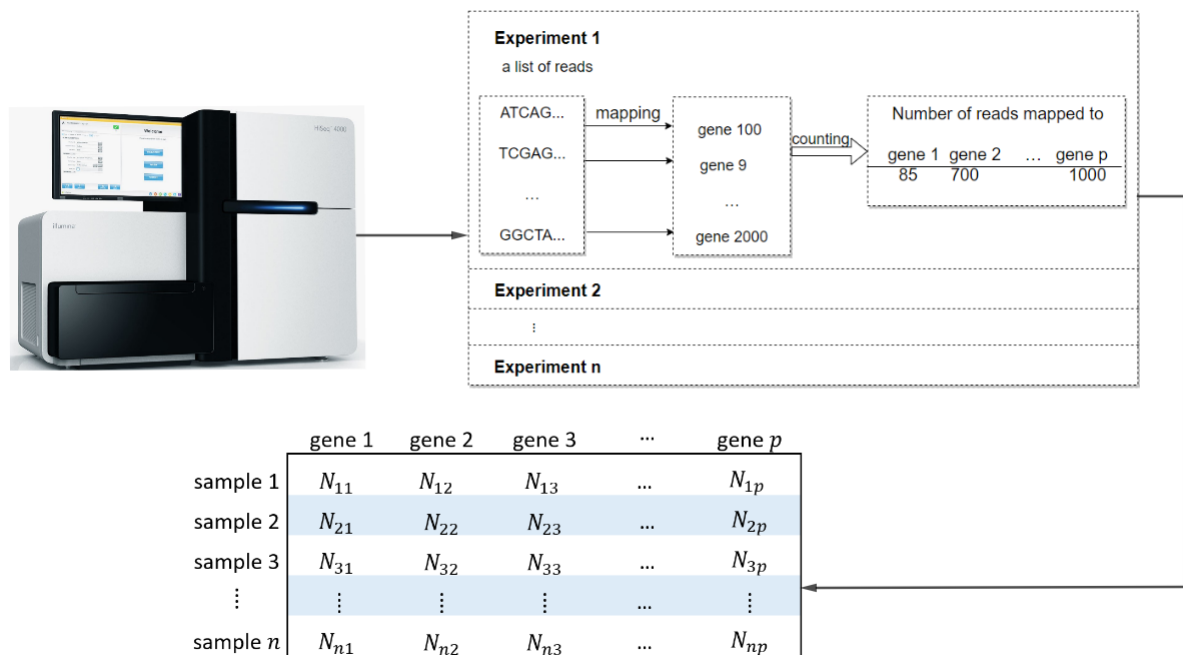


Figure 3.1: A pipeline showing how RNA–seq data matrix is obtained. Illumina HiSeq machine allows several samples to be run in parallel producing a matrix. For every sample, there exists a p number of genes. This reveals the level of expression of each gene for every sample.

3.2 Describing RNA-seq data used

This study compares RNA-seq data of 1212 samples with each having 20532 genes. To be specific, the data used for this study is [The Cancer Genome Atlas \(TCGA\)](#) which have analysed the transcriptome of 30 human tumours. In this data, 1100 patients are labelled as breast cancer patients while 112 are labelled as non-breast cancer patients. In classification tasks, data with an unequal number of instances for different classes are known to be unbalanced. The data used for this study is considered as an unbalanced data because it does not have equal instances for the breast cancer class and non-breast cancer class. Making this data publicly available has enabled researchers to provide global information on cancer and improved diagnostic methods ([Tomczak *et al.*, 2015](#)). Furthermore, the study uses TCGA because it provides corresponding clinical data for the samples. For example, follow-up time and patient status (alive or dead) are provided in the clinical data which will be used for survival analysis.

3.3 Extraction of essential genes

Significance Analysis of Microarray (SAM) is a statistical technique used to identify genes that are significantly different in an expression data. Previously, it was used only to analyse microarray experiments, but the recent RNA-seq technology led [Li and Tibshirani \(2013\)](#) to develop [SAMseq](#). The algorithm takes in gene expression data of a sample and compares with a response variable. Examples of the response variable are tumour versus normal samples, treated and untreated, level of glucose in patient's blood, and survival time of patients ([Chu *et al.*, 2001](#)). In contrast to microarray data, RNA-seq data are in the form of counts hence models based on Gaussian assumptions are not suitable because they are positive integers and can be very skewed ([Chu *et al.*, 2001](#); [Li and Tibshirani, 2013](#)).

Statistical techniques used for microarray data are mostly parametric while RNA-seq data uses non-parametric methods. For instance, taking the case of tumour versus normal patients, SAM becomes analogous to a t-test when gene expression data is from a microarray experiment and it becomes a Mann-Whitney-Wilcoxon test if it is RNA-seq ([Li and Tibshirani, 2013](#)). The focus of SAM for

this study is specifically on two-class unpaired output because the response variable has tumour and normal patients. The Mann–Whitney–Wilcoxon test statistic is given by

$$T_j = \sum_{t \in C_k} R_{tj}(N) - \frac{n_k(n+1)}{2}, \quad (3.3.1)$$

where $R_{tj}(N)$ is the rank of N_{ij} for $i = 1, \dots, n$, $j = 1, \dots, p$, n_k are sample sizes with $k = 1, 2$ for this study, and C_k is the class (tumour or normal) for the data.

To process Equation (3.3.1) correctly, RNA-seq data must be normalised by using sequencing depth. This is because samples express a different number of reads in RNA-seq data making counts N_{ij} also depend on the total number of reads generated by sample i . For example, if the total number of reads for two samples are $\sum_{j=1}^p N_{1j} = 1 \times 10^6$ and $\sum_{j=1}^p N_{2j} = 2 \times 10^6$ and their genes are not differentially expressed, then probably $2N_{1j} \simeq N_{2j}$ for any $j = 1, \dots, p$. In such a situation, we say sample 2 is relatively expressing twice as much as sample 1. Thus to compare different samples, their sequencing depths must be the same. To overcome this problem, SAMseq uses a resampling technique to short-list samples with a probability distribution as follows

$$N'_{ij} \sim \text{Poisson}\left(\frac{\bar{d}}{d_i} N_{ij}\right), \quad (3.3.2)$$

where \bar{d} is a geometric mean of the sequencing depths defined as $\bar{d} = \left(\prod_{i=1}^n d_i\right)^{\frac{1}{n}}$. d_i denotes sequencing depth and is estimated by iterating a two-step function. Let $N_{i\cdot} = \sum_{j=1}^p N_{ij}$, $N_{\cdot j} = \sum_{i=1}^n N_{ij}$, $N_{\cdot\cdot} = \sum_{i=1}^n \sum_{j=1}^p N_{ij}$, then the first step of estimating sequencing depth d_i , is given by

$$\hat{d}_i = \frac{\sum_{j \in S} N_{ij}}{\sum_{j \in S} N_{\cdot j}}, \quad (3.3.3)$$

where S is a set of genes that are not differentially expressed.

The second step involves the prediction of the set of genes, S , using the goodness-of-fit (GOF) statistic given by

$$\text{GOF}_j = \sum_{i=1}^n \frac{(N_{ij} - \hat{d}_i N_{\cdot j})^2}{\hat{d}_i N_{\cdot j}}, \quad (3.3.4)$$

and genes in S are those with GOF_j values in the $(\epsilon, 1 - \epsilon)$ quantile of all GOF_j values, where $\epsilon \in (0, \frac{1}{2})$ is a fixed constant. The predicted set S is used in Equation (3.3.3) to update \hat{d}_i which is used in Equation (3.3.4) to update GOF_j until \hat{d} becomes invariant. \hat{d}_i is initially set to $\frac{N_i}{N}$ (Li *et al.*, 2012).

After resampling, Equation (3.3.1) changes to

$$T'_j = \sum_{t \in C_k} R_{tj}(N') - \frac{n_k(n+1)}{2}. \quad (3.3.5)$$

However, ties may exist in $R_{tj}(N')$ in Equation (3.3.5) hence, small random numbers are added to each count. This results in $N'_{ij} \leftarrow N'_{ij} + \epsilon_{ij}$ where $\epsilon_{ij} \sim \text{Uniform}(0, 0.1)$ for $1 \leq i \leq n, 1 \leq j \leq p$. Unfortunately, the introduction of ϵ_{ij} decreases the power of Equation (3.3.5) by increasing its variance. Also, a significant amount of reads are not included during the resampling technique. To increase the power of Equation (3.3.5), the geometric average \bar{d} of a repeated resampling technique in Equation (3.3.2) is computed. This gives the statistic

$$T_j^* = \frac{1}{L} \sum_{l=1}^L \left(\sum_{t \in C_k} R_{tj}(N^l) - \frac{n_k(n+1)}{2} \right), \quad (3.3.6)$$

where L is the number of times resampling is repeated for $l = 1, \dots, L$ and N^l represents the l th N' resampling.

In brief, SAM computes T_j^* for each gene by first estimating sequencing depth and applying resampling technique using the estimated sequencing depth. Thereafter, the Mann–Whitney test statistic is applied to each resampled data to obtain a summary measure. Finally, an average of the summary measure is then computed. Below are the steps involved in SAMseq;

- i. Compute T_1^*, \dots, T_p^* for $j = 1, \dots, p$.
- ii. Order statistic such that $T_{(1)}^* \leq T_{(2)}^* \leq \dots \leq T_{(p)}^*$.
- iii. Permute the response values y_j B times. For each permutation b , compute statistic T_j^{*b} and their corresponding order $T_{(1)}^{*b} \leq T_{(2)}^{*b} \leq \dots \leq T_{(p)}^{*b}$.
- iv. Estimate the expected order statistic by $\bar{T}_{(j)}^* = \frac{1}{B} \sum_b T_{(j)}^{*b}$ from the set of B permutations.

- v. Plot $T_{(j)}^*$ values versus the $\bar{T}_{(j)}^*$.
- vi. We then choose a fixed threshold (Δ), such that all genes with $|T_{(j)}^*| > \Delta$ are called significant genes.

Depending on the gene expression and the class of the sample, T_j^* may either be positive or negative. A gene in which higher expression correlates with higher values of the class is known as significant positive genes. Significant positive genes can also be lower expression correlating with the lower values of the class. For example, in an experiment with two classes labelled 1 and 2, genes with higher expression correlating with class 2 are significant positive genes. Similarly, a significant negative gene has lower expression correlating with higher values of the class.

Even though genes with $|T| > \Delta$ are considered significant, some are just identified by chance. It is therefore important to measure the accuracy of the significant genes. In this case, the False Discovery Rate (FDR) is used to compute for the rate at which genes are identified by chance. It is defined as the expected proportion of false positive in the significant genes. Hence, FDR is estimated by

$$\text{FDR}_\Delta = \frac{\hat{\pi}_0 \hat{V}}{\hat{R}}, \quad (3.3.7)$$

where

- $\hat{\pi}_0 = 2 \sum_{j=1}^p I_{(|T_j^*| \leq q)}$; q is the median of all permuted values $|T_j^{*b}|$, and I represents an indicator function,
- $\hat{V} = \frac{1}{B} \sum_{j=1}^p \sum_{b=1}^B I_{(|T_j^{*b}| > \Delta)}$ and it represents the number of false positives in the significant genes. The numerator in Equation (3.3.7) gives the median number of falsely called genes,
- $\hat{R} = \sum_{j=1}^p I_{(|T_j^*| > \Delta)}$ represents the total number of genes that are significant.

The study also seeks to use the differential gene expression identified to classify breast cancer. After genes have been found to be differentially expressed, machine learning techniques are applied to build a model to classify breast cancer.

3.4 Exploring different classification algorithms

Machine learning are techniques which enable computers to learn from data with the help of statistical techniques. Specifically, it is used to identify patterns and hidden insights by exploring relationships in data which is then used to build models for prediction. It has two main types and they are supervised learning and unsupervised learning. A study which seeks to perform classification tasks with labelled data applies supervised learning technique. For example, this study uses gene expression data of breast cancer to classify patients by employing machine learning techniques. This is achieved by using knowledge obtained from labelled data. Examples include random forests and support vector machines. In a case where data is unlabelled, unsupervised learning is employed. The latter technique, on the other hand, has no prior knowledge but makes decisions by clustering patients based on the closeness of their gene expression. Evidently, unsupervised techniques cluster by using distant measures. Examples include k-means and hierarchical clustering (Brown *et al.*, 2000). The data used for this study is labelled, thus, this section presents three supervised machine learning techniques to be used for the study.

3.4.1 Explaining random forests

Random forest is a supervised machine learning algorithm of ensemble decision trees that works by using bootstrap aggregate (bagging) to decorrelate features to perform classification or regression tasks (Breiman, 2001). Bootstrap however simply means to sample with replacement. Hence, bagging is a technique that bootstraps instances to create trees and aggregate them, which, as a result, reduces variance in the prediction function (Efron and Hastie, 2016). Breiman (2001) then introduced decorrelation of features to add an additional layer of randomness to bagging. An example of a tree is shown in Figure 3.2. A decision tree can be used for a classification or a regression task.

Classification and regression trees both build models by recursively partitioning the training data in a feature space to achieve maximum pureness (Strobl *et al.*, 2009). Pureness is defined by a node containing only a single class whereas impureness is defined as a node containing several classes. At each stage, the algorithm utilizes the concept of information gain (IG) to split variables. The best

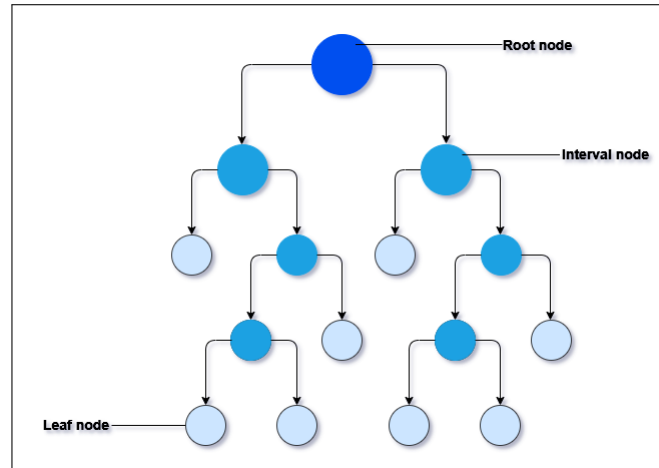


Figure 3.2: A diagram showing an example of a decision tree. The first topmost node is called the **root node** and only have arrows leaving it. Nodes having arrows pointing to it and arrows leaving it are known as **interval node**. **Leaf nodes** only have arrows pointing to it.

splitting variable is the one with the maximum IG. In other words, IG is used to determine the impurity before splitting and the impurity after splitting. This is done to obtain the level of impureness among the variables.

Given training data: $D = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$ with x_i equal to a vector of predictor variables and y_i is response, for $i = 1, 2, \dots, n$, IG is defined by

$$\text{IG} = \underbrace{\text{Impurity}(Z)}_{\text{Impurity before split}} - \underbrace{\sum_{j=1}^k \frac{|Z_j|}{|Z|} \text{Impurity}(Z_j)}_{\text{Impurity after split}}, \quad (3.4.1)$$

where $Z \subseteq D$ and is a subset of instances from the data, $|Z|$ is the total number of instances in Z , Z_j are instances that belong to class j for $j = 1, \dots, k$ and $|Z_j|$ is the number of instances that belong to class j . Since classification and regression deal with discrete and continuous dependent variables respectively, they have different impurity measures. Regression trees employ mean square error as their impurity measure. It is given by

$$\text{Impurity}(Z) = \frac{1}{|Z|} \sum_{i=1}^{|Z|} (f(x_i) - y_i)^2. \quad (3.4.2)$$

Unlike regression, there are three different impurity measures that classification trees can apply. These are misclassification error, Gini index, and cross entropy. Let p_j be the probability of class j , then details of the three impurity measures are given below:

i. The misclassification error is given by

$$\text{misclassification error} = 1 - \max\{p_j\}. \quad (3.4.3)$$

The misclassification error of a pure variable is zero (0) and its value is always in a closed unit interval $[0, 1]$. This is because, for a pure variable, the probability is 1 therefore

$$1 - \max\{1\} = 1 - 1 = 0.$$

ii. The Gini index is given by

$$\text{Gini Index} = 1 - \sum_{j=1}^k p_j^2. \quad (3.4.4)$$

Gini index of a pure variable is zero (0) because, for a pure variable, the sum of the probabilities is 1 therefore

$$1 - 1^2 = 1 - 1 = 0.$$

The values of the Gini index is also in a closed unit interval $[0, 1]$. Gini index reaches its maximum value when all the classes of a variable have the same probability. The maximum value of the Gini index is the same as the maximum value for the misclassification error (Teknomo, 2009). Assuming we set an equal probability of $\frac{1}{v}$ for v classes, then the maximum value for the Gini index is

$$1 - \sum_{j=1}^v \left(\frac{1}{v}\right)^2 = 1 - v \left(\frac{1}{v}\right)^2 = 1 - \frac{1}{v}.$$

and maximum value for misclassification error is

$$1 - \max\left\{\frac{1}{v}\right\} = 1 - \frac{1}{v}.$$

iii. The cross entropy is given by

$$\text{cross entropy} = - \sum_{j=1}^k p_j \log_2 p_j. \quad (3.4.5)$$

Cross entropy of a pure variable is also zero (0) because $1 * \log_2(1) = 0$. Also, the cross entropy reaches its maximum value when all the classes of a variable have the same probability. Again, assuming there are v classes having the same probability $\frac{1}{v}$, then the maximum value of cross entropy is

$$-\sum_{j=1}^v \left(\frac{1}{v}\right) \log_2 \left(\frac{1}{v}\right) = -v \left(\frac{1}{v}\right) \log_2 \left(\frac{1}{v}\right) = -\log_2 \left(\frac{1}{v}\right) = \log_2 v.$$

This process of partition continues until a stopping criterion is reached. Examples of stopping criteria may be a maximum number of objects in a leaf node, when IG becomes invariant, setting a maximum depth for nodes.

In general, ensemble trees perform better than single trees (Liaw and Wiener, 2002). Breiman (2001) therefore discovered a relationship between the upper bound of the generalization error and correlation among the individual trees. He realised that the lower the correlation between the single trees, the lower the error. This is because features that were not selected in a tree have an opportunity of being selected in another tree resulting in a decorrelated prediction function. Random forests aggregate all the trees to build one prediction model. This is achieved by averaging in regression and by majority voting in classification. In short, ensemble trees have an advantage of including every feature in different trees which after combining them can lead to obtaining significant effects on the response (Strobl *et al.*, 2009). Below is the algorithm for random forests.

Algorithm 1: Random Forest algorithm

Data: training set $D = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$

Result: Ensemble $(T_1(x), \dots, T_B(x))$

- 1 initialization;
 - 2 B: number of iterations;
 - 3 **for** $b = 1, \dots, B$ **do**
 - 4 Draw a bootstrap sample of D^* of size n from training data;
 - 5 Select m variables at random from the p variables;
 - 6 Grow a random forest (decorrelated) tree T_b ;
-

After getting the ensemble of decision trees, a single prediction model is built by;

- i. Regression: $f(x) = \frac{1}{B} \sum_{b=1}^B T_b(x)$.
- ii. Classification: majority vote of all decision trees predictions $T_b(x), b = 1, \dots, B$.

In addition, the random forest has only two parameters. They are the number of features for each split which is denoted by m and the number of trees also denoted by B . Typically, the number of splitting features for classification is computed as \sqrt{m} , whereas for regression is given as $\frac{m}{3}$. Training samples that were discarded during the bootstrap sampling are known as Out-of-bag (OOB). OOB data are therefore used as validation data to test the prediction function. Hence, the OOB can be used to tune the train model to obtain optimal results. Specifically, OOB error obtained from the validation is used to determine the optimal number of trees for the model.

Another important type of information that random forests provide is variable importance. Random forests use OOB data to identify variables that are important to the model. When trees are constructed, the OOB data are used for validation and its prediction accuracy is recorded. During the validating process, values in the variables are permuted and its prediction accuracy computed. This results in a decrease in prediction accuracy. These accuracies are averaged over all the trees constructed, to determine the prediction strength of each variable (Friedman *et al.*, 2001).

3.4.2 Explaining artificial neural network (ANN)

An Artificial Neural Network (ANN) is a mathematical model designed to simulate the function and structure of the brain. The motivation behind this is to build a model that is able to perform human-related tasks. Humans are mainly able to perform several tasks because of the brain and it has about 86 million neurons. Generally, the neurons receive signals through dendrites which undergo a mechanism to produce an output in the axon terminal. The output from the neuron multiplicatively interacts with synapses which may become the input for other neurons. Mathematically, input data is passed through a neuron of a layer (a layer is a collection of neurons which operate together in a specific depth) where it multiplicatively interacts with weights. The weight corresponds

to the strength of connected neurons. To get the desired output from the interaction, an activation function is applied to generate an output which then may become an input for another neuron in a different layer (Karpathy, 2017). The process where the output of one layer becomes the input for the next layer without loops is termed as feedforward (Raschka, 2015). Finally, an output of an ANN is a function of learned weights and bias. The network learns by using gradient descent to continuously update the weight and bias. This process is combined with a chain rule technique called backpropagation. Figure 3.3 depicts how the neuron of a brain is mathematically modelled to ANN.

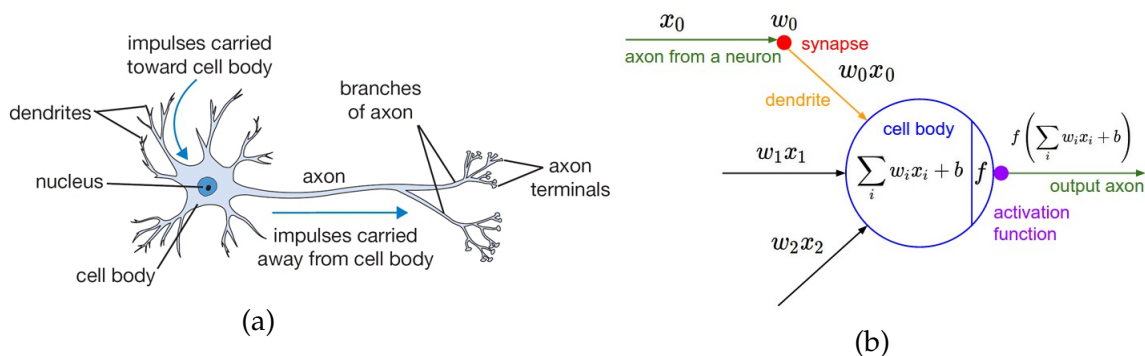


Figure 3.3: The analogy of biological neuron 3.3a used to mathematically model artificial neural network 3.3b. Both have a cell body which receives signals through the dendrite to the output (Karpathy, 2017).

An ANN is mainly characterised by three different layers and they are input layer, hidden layer, and an output layer. The first and last layers of a neural network are the input layer and output layer respectively. The hidden layer is found between the input and output layer. Also, each layer contains neurons and these are the smallest unit of a neural network. The neurons from a layer are only able to connect to the neurons of an immediate layer. They are connected by weights and is achieved by computing for the weighted sum of the neurons with the current layer. Figure 3.4 shows a simple 3-layered neural network with only one output.

The explanations above can be mathematically expressed as

$$a^{(L)} = f(z^{(L)}), \quad (3.4.6)$$

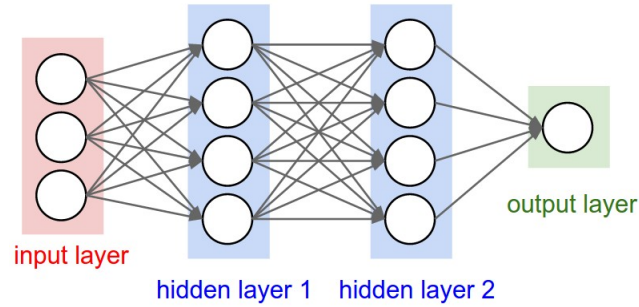


Figure 3.4: Design of a 3-layered artificial neural network. Specifically, this neural network has one input layer, two hidden layers and one output layer. Each layer has neurons and they are the smallest unit of a neural network. The layers are also connected by weights (Karpathy, 2017).

where L denotes layer, $a^{(L)}$ is the output of the current layer, f is an activation function and $z^{(L)}$ is given by

$$z^{(L)} = \mathbf{w}^{(L)}a^{(L-1)} + b^{(L)},$$

where $\mathbf{w}^{(L)}$ is the weights for the current layer, $a^{(L-1)}$ is the input for the current layer which is obtained from the output of the previous layer. $b^{(L)}$ also represents the bias of the current layer. In brief, an ANN computes for the summation of the dot product of input and weight to obtain results that are linear. This is followed by an activation function on the results to obtain the non-linear output to be fed to different layers.

An ANN without an activation function may simply be a linear function. Unfortunately, linear functions are limited in learning from complex data, such as images, videos, audio and data with higher dimensions. Evidently, activation functions are relevant for ANN because they introduce non-linear properties in the network which helps it to learn from complex data. In addition, an activation function should be differentiable so as to perform backpropagation. Popular activation functions used include the sigmoid function, hyperbolic tangent function, and rectified linear unit (ReLU).

- i. The sigmoid function is denoted by $\sigma(z)$ and is expressed as

$$\sigma(z) = \frac{1}{1 + \exp(-z)}.$$

This activation function outputs values ranging between 0 and 1. Specifically, large negative outputs tend to have numbers closer to 0 whereas large

positive outputs tend to have numbers closer to 1. Unfortunately, when a sigmoid function is applied, most neurons may get values close to either 0 or 1. When this happens, the gradient of the neurons vanishes as weights and biases are learned. This eventually makes the network slow to learn and yield poor accuracies.

- ii. The hyperbolic tangent function $\tanh(x)$ is expressed as

$$\tanh(z) = \frac{1 - \exp(-2z)}{1 + \exp(-2z)}.$$

The hyperbolic tangent function also outputs values ranging between -1 and 1. Similar to the sigmoid function, large negative outputs tend to have numbers closer to -1 while large positive outputs tend to have numbers closer to 1. Hyperbolic tangent functions also suffer from vanishing gradient problems.

- iii. The Rectified Linear Unit (ReLU) is given by

$$f(z) = \max\{0, z\}.$$

ReLU simply takes in input signals and return zero if it is negative but returns the same number if it is positive. Unlike sigmoid and hyperbolic tangent function, ReLU does not suffer from vanishing gradient problems and is not computationally expensive. ReLU transforms linear results into non-linear outputs, but still remain close to the linear space. It has the ability to preserve many properties that make linear models easy to optimize with gradient-based methods, therefore it is considered as nearly linear. Thus for feedforward neural networks, the recommended activation function is ReLU (Goodfellow *et al.*, 2016).

In addition, this study seeks to classify breast cancer samples. Again, the data used has two classes; namely tumour or tumour-free. For this reason, the output layer will also need another activation function called the softmax function. The output layer uses the softmax function because the desired results must be in the form of either 0 or 1. For K -dimensional data, the softmax function is given by

$$\sigma(z_j) = \frac{\exp(z_j)}{\sum_{k=1}^K \exp(z_k)}. \quad (3.4.7)$$

As it can be seen in Equation (3.4.7), the outputs of softmax function are probabilities, therefore their sum is equal to 1. Although the sigmoid function also ranges between 0 and 1, the softmax function is preferred because it is able to take in a K -dimensional vector of real numbers and return a K -dimensional vector of probabilities.

It is also important to measure the performance of the network. This is achieved by measuring the discrepancy between the target value and the output produced by the network (LeCun *et al.*, 2012). The function used to check performance is known as a cost function. For a classification problem, the cost function computes the average of cross entropy for all m training examples.

Let y_i be the target value for training example i and if $a^{(L)}$ is the output layer, then the cost function is given by

$$\mathcal{J}(\mathbf{w}, b) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(y_i, a_i^{(L)}),$$

where $\mathcal{L}(y_i, a_i^{(L)})$ is cross entropy defined as

$$\mathcal{L}(y_i, a_i^{(L)}) = - \sum_{j=1}^K y_i \log(a_i^{(L)}).$$

Next, the network seeks to minimize the cost function by going through the backpropagation process. Backpropagation uses the gradient to gradually optimize the weights of the network. This is important because it helps in understanding how sensitive the cost function is to small changes in the weights. In others words, backpropagation gives the derivative of the cost function with respect to weight. This is made possible by using the chain rule to iteratively compute gradients for each layer. It is hence computed as follows:

$$\frac{\partial \mathcal{J}(\mathbf{w}, b)}{\partial \mathbf{w}^{(L)}} = \frac{\partial z^{(L)}}{\partial \mathbf{w}^{(L)}} \frac{\partial a^{(L)}}{\partial z^{(L)}} \frac{\partial \mathcal{J}(\mathbf{w}, b)}{\partial a^{(L)}}. \quad (3.4.8)$$

The weights are optimized using an optimization algorithm called gradient descent. It is defined as

$$\mathbf{w} := \mathbf{w} - \eta \frac{\partial \mathcal{J}(\mathbf{w}, b)}{\partial \mathbf{w}^{(L)}}, \quad (3.4.9)$$

where η is a learning rate. It is a hyper-parameter that determines the rate at which the weights are updated. High learning rate means the weights will be updated at a faster rate. The problem with high learning rate is, the parameter may never settle at the minima causing the cost function to fluctuate or diverge. Small learning rate also means the weights will be updated at a slower rate. The problem with small learning rate is, the weights take long to learn. Figure 3.5 displays the effect of different learning rates on the loss function.

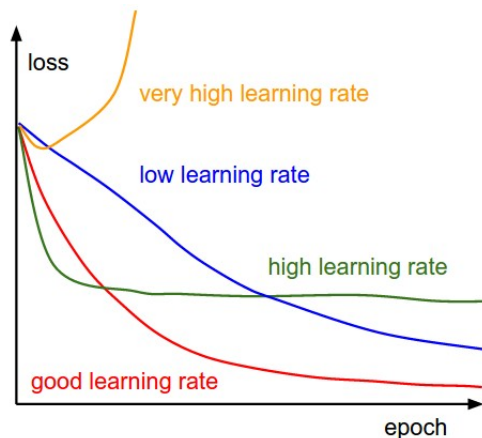


Figure 3.5: A diagram showing the effect of different learning rate on loss function (Karpathy, 2017).

The line labelled as *very high learning rate* has exploded and it may never settle at a minima. This suggests that a gradient descent with a large learning rate may cause the loss function to diverge. Also, the line labelled as *high learning rate* initially was fast to approaching the loss but could not converge to the optimal loss. This suggests that a gradient descent with a high learning rate may appear to be fast initially but will be stuck after training for a while. The line labelled as *low learning rate* was gentle throughout the training process. This suggests that a gradient descent with a low learning rate will find the minima at a very slow rate. Finally, the line labelled *good learning rate* decays at a good rate and it is able to approach the loss.

Figure 3.6 shows an example of parameter finding the optimal loss in an ANN. The main aim of gradient descent is to reduce the cost function and possibly find the optimal loss. This is because, in a cost function, there may exist many local minima. In the figure, there exist only two minima. The parameter starts off

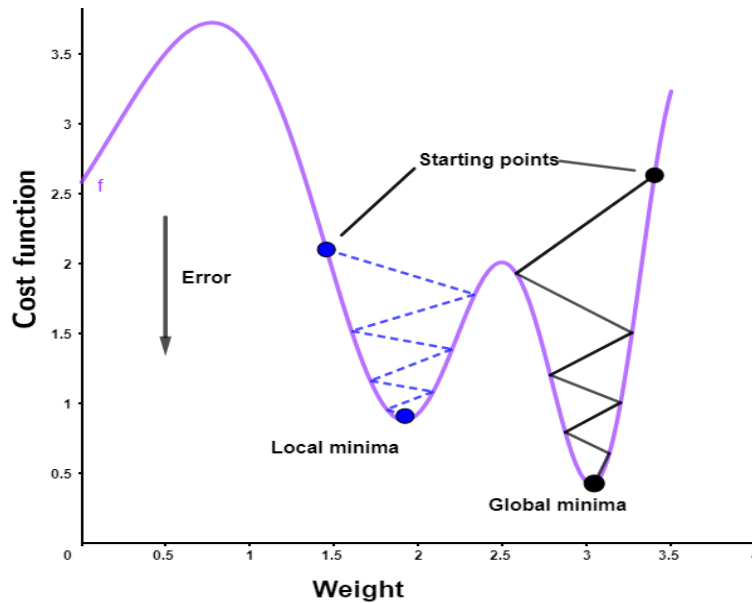


Figure 3.6: A diagram showing a cost function with two minima. The blue and black parameter starts off at different positions. Only the black parameter reaches the global minima.

by finding the steepest direction downwards until it reaches the minima. It is seen that the blue parameter starts at a steep descent and ends at a local minima while the black parameter starts at another steep descent and ends at the global minima. In such a scenario, the parameter at the global minima is preferred.

However, it should be noted that Equation (3.4.8) and (3.4.9) are only computed with respect to the weights. The weight (\mathbf{w}) is substituted with bias (b) to also determine how sensitive the cost function is to small changes in the bias. The goal of backpropagation is to train a multi-layered neural network such that it can learn the appropriate internal representations to help learn any arbitrary mapping of input to output (Rumelhart *et al.*, 1986).

In short, an ANN goes through two main phase cycles and they are feedforward and backpropagation. When a neural network is fed with an input vector, it propagates forward through different layers until it reaches the output layer. A cost function is then used to measure the performance of the network by comparing the output with the target values. The output layer is always in a K -dimensional vector which represents the number of classes for the target. Each neuron in the output layer represents a class and each has an error value. These values are therefore used to backpropagate from the output layer through

each neuron in the network to help determine their contribution to the output.

Finally, the gradient of the cost function is computed using the error values. The resulting gradient is then fed into an optimization algorithm to update the weights and bias. This process continues until the cost function reaches local or global minima.

3.4.3 Support vector machines (SVM)

Support vector machines (SVM) is a supervised machine learning algorithm which uses the concept of hyperplanes and margins to separate classes. Figure 3.7 depicts the use of a hyperplane to discriminate two groups.

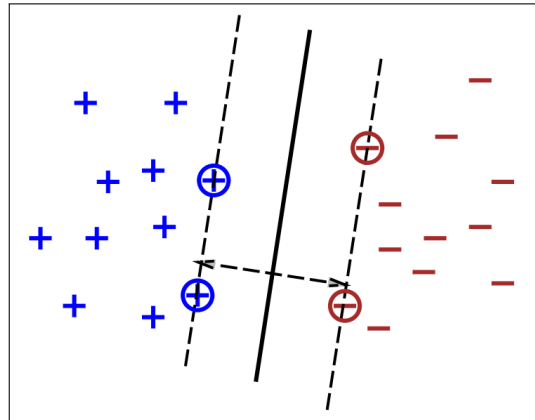


Figure 3.7: An example of a hyperplane separating the positive group from the negative group (Bottou and Lin, 2007).

The thick solid line in Figure 3.7 is called the hyperplane while the dashed lines are the margins. There may be several hyperplanes but the best is chosen based on the one that maximizes the distance between the margin (Bishop, 2016). Thus, the main objective of SVM is to optimally find a hyperplane that widens the distance between the margins to distinctly separate groups involved. It, therefore, applies an implicit function which incorporates the use of kernels for data whose classes are linearly non-separable (Friedman *et al.*, 2001). The objective function of SVM is therefore given by

$$y = \mathbf{w}^T \Phi(\mathbf{x}) + b, \quad (3.4.10)$$

where \mathbf{w} is the weight of the objective function and is also given by

$$\mathbf{w} = \sum_{i=1}^n \alpha_i t_i \Phi(\mathbf{x}_i^T), \quad (3.4.11)$$

with α_i being a Lagrange multiplier and t_i an indicator which takes a value of 1 if \mathbf{x}_i is positive and -1 if \mathbf{x}_i is negative. $\Phi(\mathbf{x}_i)$ is also a fixed feature space of \mathbf{x}_i

Equation (3.4.10) therefore becomes

$$y = \sum_{i=1}^n \alpha_i t_i \Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_i^T) + b. \quad (3.4.12)$$

Cortes and Vapnik (1995) then represented $\Phi(\mathbf{x}) \cdot \Phi(\mathbf{x}_i^T)$ with $K(\mathbf{x}, \mathbf{x}_i^T)$ which is known as the kernel function. Thus, Equation (3.4.12) is written as

$$y = \sum_{i=1}^n \alpha_i t_i K(\mathbf{x}, \mathbf{x}_i^T) + b. \quad (3.4.13)$$

Again, Cortes and Vapnik (1995) noted that SVM is a *Universal Machine* because the kernel function can take different forms leading to the implementation of networks with several functions. Hence, different kernel functions can be used to obtain an optimal hyperplane. It enables the input data to be transformed into a higher feature space \mathcal{F} , which aid in obtaining the optimal hyperplane (Tong and Koller, 2000). Kernel functions used for this study includes polynomial of order d , radial basis, sigmoid and linear function (Jean-Philippe Vert, 2001).

- i. Kernel function for a polynomial of order d is given by

$$K(\mathbf{x}, \mathbf{x}_i^T) = (\mathbf{x} \cdot \mathbf{x}_i^T + 1)^d.$$

- ii. Kernel function for radial basis is also written as

$$K(\mathbf{x}, \mathbf{x}_i^T) = \exp\left(-\frac{|\mathbf{x} - \mathbf{x}_i^T|^2}{\sigma^2}\right).$$

- iii. Kernel function for sigmoid is given by

$$K(\mathbf{x}, \mathbf{x}_i^T) = \tanh(\kappa \mathbf{x} \cdot \mathbf{x}_i^T + \theta),$$

where κ is called the gain parameter and θ is the threshold.

iv. Kernel function for linear is also given by

$$K(\mathbf{x}, \mathbf{x}_i^T) = \mathbf{x} \cdot \mathbf{x}_i^T.$$

A relevant property of the SVM algorithm is its ability to correspond to the convex space. Due to this, any local optimum will be a global solution (Bishop, 2016).

3.4.4 Median–supplement: a balancing data technique

As mention in Section 3.2, it is evident that the data is unbalanced. An unbalanced data refers to classification problems where there are unequal instances for different classes. A major problem of unbalanced data is that the model is likely to be biased to the dominant class during prediction. A popular solution to this problem is to either oversample or under–sample the data. In the case of under–sampling with two classes, a random subset of samples of the larger class is selected to match the number of samples of the smaller class. Unfortunately, there is a possibility of losing important information from the discarded samples in the larger class. For oversampling with two classes, instances in the smaller class are randomly duplicated to match the number of samples of the larger class. An advantage of oversampling is that it has a low possibility of losing relevant information but there is also a risk of overfitting because it may be more likely to obtain similar samples in the data (Glander, 2018).

A method called median–supplement was therefore introduced by Adabor and Acquaah-Mensah (2017) to balance the data. This method of balancing data uses the idea of oversampling by generating a matrix \mathbf{J} with dimension $m \times n$. m is the difference between the number of samples for the two classes while n is the number of variables for the data. The matrix is generated by first randomly generating an $m \times n$ matrix \mathbf{L} using Latin hypercube sampling with uniformly distributed values between 0 and 1 (McKay *et al.*, 1979; Stein, 1987). The procedure for generating Latin hypercube samples is described below.

Let $\mathbf{P} = (p_{jk})$ be a matrix with dimension $m \times n$, where each column of \mathbf{P} is an independent random permutation of $\{1, 2, \dots, m\}$. Also, let ξ_{jk} be $m \times n$ iid $U[0, 1]$, $j = 1, \dots, m$; $k = 1, \dots, n$. Then X_{jk} is an element in the Latin hyper-

cube matrix defined by

$$X_{jk} = F_k^{-1}(m^{-1}(p_{jk} - 1 + \zeta_{jk})), \quad (3.4.14)$$

where F is a cumulative distribution function (CDF). p_{jk} determines which cell X_j belongs to and ζ_{jk} determines where in the cell is X_j . Figure 3.8 shows Latin hypercube samples

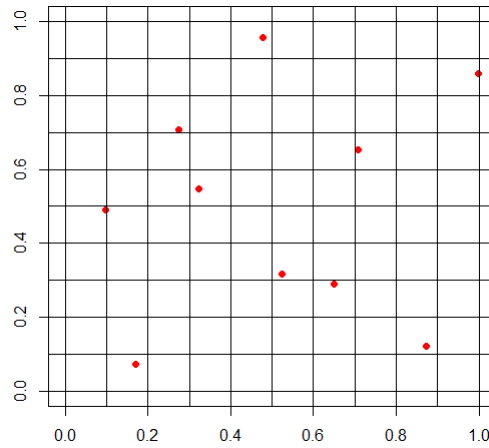


Figure 3.8: A Latin hypercube sampling with $m = 10$ and $n = 2$. Every row or column does not contain more than one point.

Figure 3.8 has been constructed in a way that each point has been located to a particular cell such that a row or column do not contain more than one point. The Latin hypercube sampling method attempts to distribute samples evenly over the sample space. It can also be seen that every point sampled is between 0 and 1. This approach of generating random samples is therefore used to construct the matrix \mathbf{L} .

Afterwards, the matrix \mathbf{L} is multiplied by the median of each variable yielding the matrix \mathbf{J} . The matrix \mathbf{J} then becomes a supplement for the data. For instance, consider data with 50 patients each with 5 attributes. If 35 of them are labelled as having breast cancer and 15 of them do not have breast cancer, this data is deemed as unbalanced. It can be balanced by generating an extra 20 instances using the median–supplement method.

The median–supplement \mathbf{J} is therefore computed by,

$$\mathbf{J}_{m \times n} = \mathbf{L} \mathbf{M}, \quad (3.4.15)$$

where \mathbf{L} is an $m \times n$ dimension matrix generated using Latin hypercube sampling with uniformly distributed values between 0 and 1. It is therefore given by

$$\mathbf{L} = \begin{bmatrix} \mathbf{0} - \mathbf{1} \end{bmatrix}_{m \times n} .$$

Also, \mathbf{M} is a $m \times n$ diagonal matrix of the median. Let the median for variable i be denoted by $\tilde{\mu}_i$, then \mathbf{M} is given by

$$\mathbf{M} = \begin{bmatrix} \tilde{\mu}_1 & & & \\ & \tilde{\mu}_2 & & \mathbf{0} \\ & & \ddots & \\ & \mathbf{0} & & \tilde{\mu}_{n-1} \\ & & & & \tilde{\mu}_n \end{bmatrix}_{m \times n} .$$

After [Adabor and Acquah-Mensah \(2017\)](#) applied this method in random forests and Naïve Bayes, they had higher accuracies in classifying breast cancer patients into their HER2 receptor status phenotype.

3.5 Performance metrics for classification algorithms

The performance metric used for this study is the confusion matrix. It is a measure used to evaluate the performance of classifiers. A confusion matrix for a two-class data can also be described as a table with four different combinations of predicted and actual values. They are true positive (TP), true negative (TN), false positive (FP), and false negative (FN). Suppose confusion matrix is denoted by C such that $C_{i,j}$ is the number of individuals observed in class i but predicted to be in class j , then in a binary classification problem, true positive is $C_{1,1}$, true negative is $C_{0,0}$, false positive is $C_{0,1}$ and false negative is $C_{1,0}$.

		Actual Values	
		Positive (1)	Negative (0)
Predicted Values	Positive (1)	TP	FP
	Negative (0)	FN	TN

Figure 3.9: A confusion matrix with two classes; positive class and a negative class. The row of the matrix represents class predicted by the classifier. The column of the matrix represents the actual class from the data.

These combinations are used to compute for sensitivity, specificity, false positive rate and false negative rate.

- i. **Sensitivity:** This is used to measure the proportion of correctly identifying actual positives. It is computed as

$$\text{Sensitivity} = \frac{TP}{TP + FN}. \quad (3.5.1)$$

For example in a breast cancer study, sensitivity is the percentage of patients who are correctly classified as having breast cancer. For sensitivity, the bigger the percentage, the better.

- ii. **Specificity:** This is used to measure the proportion of correctly identifying actual negatives. It is also computed as

$$\text{Specificity} = \frac{TN}{TN + FP}. \quad (3.5.2)$$

Again in a breast cancer study, specificity is the percentage of patients who are correctly classified to not having breast cancer. Also for specificity, the bigger the percentage, the better.

- iii. **False Positive Rate:** This is used to measure the proportion of all negatives that are falsely classified as positives. It is therefore computed as

$$\text{False Positive Rate} = \frac{FP}{FP + TN}. \quad (3.5.3)$$

Likewise in a breast cancer analysis, the false positive rate is the percentage of patients who have been wrongly classified as having breast cancer. For the false positive rate, the smaller the rate, the better.

iv. **False Negative Rate:** This is a measure used to determine the proportion of positives that are falsely classified as negatives. It is also computed as

$$\text{False Negative Rate} = \frac{\text{FN}}{\text{TP} + \text{FN}}. \quad (3.5.4)$$

Similarly, in a breast cancer study, the false negative rate is the percentage of patients who have been falsely classified as not having breast cancer. For the false negative rate, the smaller the rate, the better.

v. **Accuracy:** This is a measuring system used to determine the degree of closeness to the quantity of the true value. Hence, it is computed as

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \quad (3.5.5)$$

For example in a breast cancer study, accuracy is the percentage of correctly classifying patients to their true class. Finally, the classifier seeks to obtain the highest accuracy.

3.6 Investigating survival analysis

Survival analysis is used to understand and make inferences about time to an event. Examples of time to an event includes time to death, remission duration of a disease, failure times of a machine, completion of graduate degrees, in human lifetime data (Nasejje, 2012). Lifetime data may either be complete or incomplete. It is complete when the event of interest is observed and incomplete otherwise. Incomplete data is popularly known as censored. Generally, there are three types of censoring, namely: right censoring, left censoring and interval censoring.

Right censoring occurs when the time to an event of interest is unknown but greater than the survival time or when an event of interest happens after the study. Similarly, when time to an event of interest is less than the survival time, it is known as left censoring. Also, interval censoring happens when an event of interest is only known to be between two known survival times but the exact time is unknown. For this study, the event of interest is time to death and it only considers right censoring because death can only be observed once.

3.6.1 Defining survival function

To understand and analyse lifetime data, we use a survival function. A survival function $S(t)$ gives the probability that an object of interest will survive beyond a specified time t . In this research, the objects of interest are patients. The survival function is formally defined as

$$S(t) = P(T > t). \quad (3.6.1)$$

where T is a non-negative random variable which represents the survival time of a patient.

3.6.2 Exploring non-parametric survival models

Using the definition above, one can estimate $S(t)$ empirically as

$$\hat{S}(t) = \frac{\text{Number of individuals with survival time } \geq t}{\text{Number of individuals in the dataset}}. \quad (3.6.2)$$

Equation (3.6.2) works if no observation is censored it is limited because it does not take into account censored data. Given this difficulty, [Kaplan and Meier \(1958\)](#) used a product-limit approach to take into account the censored data. This method of estimating $S(t)$ is called Kaplan–Meier survival and is also used to generate Kaplan–Meier survival curves. To estimate the probability of surviving in an interval I_i , we have to know the number of individuals who have died in the interval.

3.6.2.1 Kaplan–Meier Survival Estimate

Let q_i be the probability of dying in the interval I_i , and \hat{q}_i be an estimate of q_i then

$$\hat{q}_i = \frac{d_i}{n_i}, \quad (3.6.3)$$

where d_i is the number of individuals or units that have died in the interval I_i and n_i is the total number of individuals at risk in the interval. We can now compute for the probability of surviving in the interval I_i .

Let p_i be the probability of surviving in the interval I_i , and let \hat{p}_i be the estimate of p_i then

$$\hat{p}_i = 1 - \hat{q}_i. \quad (3.6.4)$$

Equation (3.6.4) is therefore re-written as

$$\begin{aligned}\hat{p}_i &= 1 - \frac{d_i}{n_i}, \\ \hat{p}_i &= \frac{n_i - d_i}{n_i}.\end{aligned}\quad (3.6.5)$$

We then compute for the product of all probability of surviving within the intervals that precede a specific time t . Hence, the survival estimate for the Kaplan–Meier is given by

$$\begin{aligned}\hat{S}(t) &= \prod_{i=1}^k p_i, \\ \hat{S}(t) &= \prod_{i=1}^k \left(\frac{n_i - d_i}{n_i} \right).\end{aligned}\quad (3.6.6)$$

Kaplan–Meier survival estimate is a non–parametric method as it does not require specific parameters and assumptions to be made about the underlying probability distribution of the survival times. There are cases where it is of interest to compare the Kaplan–Meier survival curve of groups in a study. This is done by using a test statistic called the log–rank statistic.

3.6.2.2 Log–Rank Statistic

A log–rank test is used to compare survival curves of two or more groups of individuals in a given sample. Log–rank test works like the chi-square (χ^2) test which employs the observed output and the expected output to make inferences. For example, this study is interested in whether the levels of gene expressed affects survival time. This, in particular, helps identify a significant difference between patients survival time based on the level of a gene expressed. Although log–rank test can be used for multiple groups, this study focuses on two group comparison. In a log–rank test, an inference is made to either reject or accept the null hypothesis that the two groups have a significantly equivalent survival curve.

Let n_{ij} be the number of individuals at risk in group i for $i = 1, 2$ at time t_j . Also let d_{ij} be the number of observed deaths in group i for $i = 1, 2$ at time t_j , then $O_i = \sum_j d_{ij}$ is the total number of observed deaths in group i . Similarly, let e_{ij} be

the expected deaths in group i for $i = 1, 2$ at time t_j , then $E_i = \sum_j e_{ij}$ is the total expected deaths in group i , where

$$e_{ij} = \frac{n_{ij}}{n_i} \times (d_{1j} + d_{2j}).$$

The log-rank test statistic is therefore given by

$$T_L = \frac{(O_i - E_i)^2}{\text{Var}(O_i - E_i)}, \quad (3.6.7)$$

and

$$\text{Var}(O_i - E_i) = \frac{n_{1j}n_{2j}(d_{1j} + d_{2j})(n_{1j} + n_{2j} - d_{1j} - d_{2j})}{(n_{ij} + n_{2j})^2(n_{1j} + n_{2j} - 1)}.$$

One disadvantage of using the non-parametric model is that it cannot be applied to data with multiple covariates. Also, it does not give any information on the best probability distribution that describes the data. This is the reason why parametric models are needed.

3.6.3 Exploring specific probability distributions

Generally, the death of an individual in a study may be caused by several factors and can be difficult to estimate mathematically. Applying theoretical distribution can make things easier in describing survival data. In this section, we explore common theoretical distributions used to analyse survival time and their applications. These theoretical distributions are known as probability distributions. They are defined by a finite number of parameters and underlying assumptions. Such parameters include the scale parameter, shape parameter and location parameter.

A scale parameter expresses how samples are spread in a distribution. Examples include standard deviation and variance. Location parameter is used to determine the shift of the distributions. Popular examples of the location parameter are mean, median and mode. A shape parameter determines the shape of the distribution and an example is skewness.

In probability distributions, cumulative distributive function (CDF) gives the

probability that a continuous random variable T (an object of interest) will have a value less than t . A CDF is denoted by $F(t)$ and is given by

$$F(t) = P(T \leq t). \quad (3.6.8)$$

This definition creates a complementary relationship between $S(t)$ and $F(t)$. Equation (3.6.1) can therefore be written as

$$S(t) = P(T > t) = 1 - P(T \leq t).$$

Hence from Equation (3.6.8)

$$S(t) = 1 - F(t). \quad (3.6.9)$$

When $F(t)$ is differentiated, we get another function called probability density function (PDF). It is denoted by $f(t)$ and is expressed as

$$f(t) = \frac{dF(t)}{dt}. \quad (3.6.10)$$

This implies that the $F(t)$ can also be expressed as

$$F(t) = \int_{-\infty}^t f(u) du = P(T \leq t). \quad (3.6.11)$$

Another function used to describe the distribution of T is the hazard function. As the survival function gives the probability that an event will occur past a time t , the hazard function gives the instantaneous potential of the event occurring per unit time, given that an individual has survived up to time t . The hazard function thus gives the risk of the event occurring per unit time during the study. The hazard function is denoted by $h(t)$ and it is formally given by

$$h(t) = \lim_{dt \rightarrow 0} \frac{P(t < T \leq t + dt | T > t)}{dt}. \quad (3.6.12)$$

The numerator of Equation (3.6.12) can be interpreted as the conditional probability that an individual will experience the event of interest in the interval $(t, t + dt]$ given that he has survived beyond time t . The denominator also describes the length of the interval. Thus, the function gives an instantaneous rate of the event occurring as the limit of the interval approaches zero (Rodriguez, 2007).

Using the approach of conditional probability, we can write Equation (3.6.12) as

$$h(t) = \lim_{dt \rightarrow 0} \frac{P(t < T \leq t + dt)}{dt P(T > t)}, \quad (3.6.13)$$

but from Equation (3.6.8), we can also write

$$F(t + dt) = P(T \leq t + dt).$$

Hence we can write $h(t)$ as

$$\begin{aligned} h(t) &= \lim_{dt \rightarrow 0} \frac{F(t + dt) - F(t)}{dt S(t)}, \\ h(t) &= \lim_{dt \rightarrow 0} \frac{F(t + dt) - F(t)}{dt} \frac{1}{S(t)}, \\ h(t) &= \frac{dF(t)}{dt} \frac{1}{S(t)}. \end{aligned} \quad (3.6.14)$$

From Equation (3.6.10), we can write Equation (3.6.14) as

$$h(t) = \frac{f(t)}{S(t)}. \quad (3.6.15)$$

Relationships of functions in survival analysis

Interestingly, there exists a mathematical relationship between survival function and hazard function. This means that they can be derived from each other. From Equation (3.6.9), we can write Equation (3.6.10) as

$$\begin{aligned} f(t) &= \frac{d(1 - S(t))}{dt}, \\ f(t) &= \frac{-dS(t)}{dt}, \end{aligned} \quad (3.6.16)$$

but from Equation (3.6.15), we can write Equation (3.6.16) as

$$\begin{aligned} h(t) &= \frac{-dS(t)}{dt} \times \frac{1}{S(t)}, \\ h(t) &= \frac{-dS(t)}{S(t)} \times \frac{1}{dt}, \\ h(t) &= \frac{-d(\ln(S(t)))}{dt}, \end{aligned} \quad (3.6.17)$$

Multiplying (3.6.17) through by dt

$$h(t) d(t) = -d(\ln(S(t))). \quad (3.6.18)$$

Integrating both sides, in Equation (3.6.18),

$$\int_0^t h(u) du = -\ln(S(t)),$$

$$S(t) = \exp\left(-\int_0^t h(u) du\right). \quad (3.6.19)$$

Given that we have the hazard function, we can always take the exponent of the integrated hazard function to obtain the survival function. Thus, when the survival function is specified, the hazard function can be obtained by taking the logarithm of the differentiated survival function given by (3.6.17).

Probability distributions to be used for this study are exponential, Weibull, log-logistic, log-normal and extreme-value model. These models assume that a continuous random variable T follows a specific distribution.

3.6.3.1 Exponential distribution

An exponential distribution of a continuous random variable T is characterised by only a scale parameter $\lambda > 0$ and is written as $T \sim \text{exponential}(\lambda)$. The probability density function of the exponential distribution is given by

$$f(t) = \lambda \exp(-\lambda t), \quad t > 0.$$

The CDF for the exponential distribution is given by

$$F(t) = 1 - \exp(-\lambda t).$$

From the relation in Equation (3.6.9), we can write the survival function of the exponential distribution as follows

$$S(t) = \exp(-\lambda t).$$

From Equation (3.6.15), we can write the hazard function as

$$h(t) = \frac{\lambda \exp(-\lambda t)}{\exp(-\lambda t)} = \lambda. \quad (3.6.20)$$

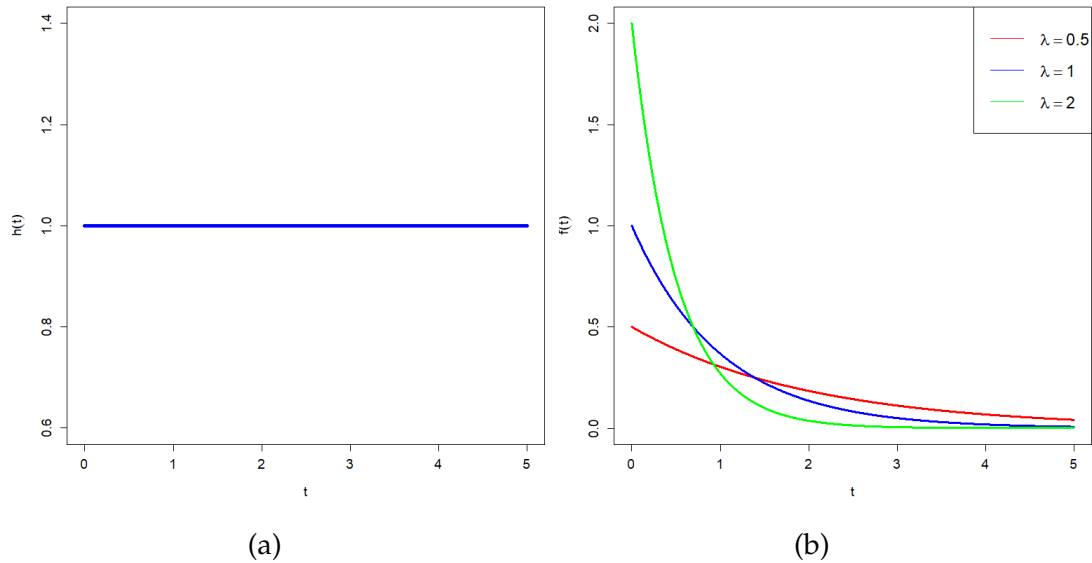


Figure 3.10: Figure 3.10a shows a plot of the hazard function $h(t)$ with $\lambda = 1$ and Figure 3.10b shows a probability density function $f(t)$ for the exponential distribution with different λ 's.

The exponential distribution is popularly known for its memoryless property. As seen from Equation (3.6.20) and Figure 3.10a, the hazard function is a constant which implies that risk does not change over time in an exponential distribution. On the whole, an individual's age does not have any effect on his survival in an exponential distribution. Furthermore, a lower λ implies a higher survival probability with a lower risk of experiencing the event and vice versa.

3.6.3.2 Weibull distribution

The Weibull distribution of a continuous random variable T is characterised by two parameters. They are the scale parameter $\alpha > 0$ and the shape parameter $\gamma > 0$. It is expressed as $T \sim \text{Weibull}(\alpha, \gamma)$. The probability density function is given by

$$f(t) = \frac{\gamma}{\alpha} t^{\gamma-1} \exp\left(-\frac{1}{\alpha} t^\gamma\right), \quad t > 0.$$

The CDF for the Weibull distribution is also given by

$$F(t) = 1 - \exp\left(-\frac{1}{\alpha} t^\gamma\right).$$

Using the relation in Equation (3.6.9) the survival function can be written as

$$S(t) = \exp\left(-\frac{1}{\alpha}t^\gamma\right).$$

The hazard function is therefore given by

$$h(t) = \frac{\frac{\gamma}{\alpha}t^{\gamma-1} \exp\left(-\frac{1}{\alpha}t^\gamma\right)}{\exp\left(-\frac{1}{\alpha}t^\gamma\right)} = \frac{\gamma}{\alpha}t^{\gamma-1}. \quad (3.6.21)$$

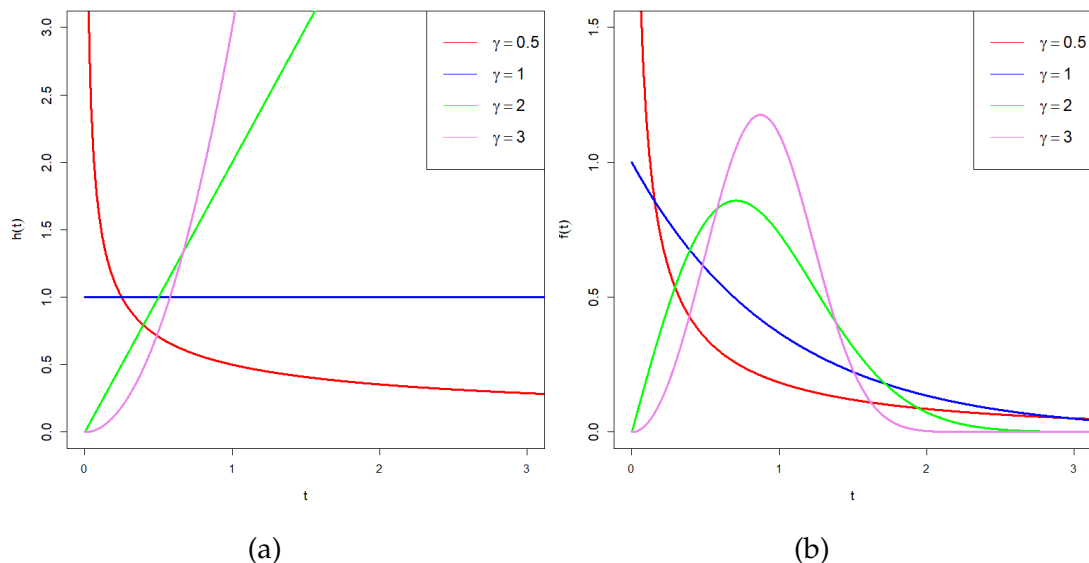


Figure 3.11: Figure 3.11a shows a plot of the hazard function $h(t)$ and Figure 3.11b shows a probability density function $f(t)$ for the Weibull distribution for different γ 's and $\alpha = 1$.

The Weibull distribution is a general form of the exponential distribution. The only difference between them is the shape parameter γ . The shape parameter is very important because it determines the behaviour of the distribution. For instance, from Figure 3.11a when $\gamma = 1$, Weibull becomes an exponential distribution. When $\gamma > 1$, the hazard rate increases as t goes to infinity and reduces when $\gamma < 1$. This makes it versatile and easier to apply to different survival problems. Lee and Wang (2003) mentioned that Weibull distribution has been applied in several survival problems and human disease mortality.

3.6.3.3 Log-logistic distribution

The log-logistic distribution is characterised by two parameters namely, scale parameter $\lambda > 0$ and shape parameter $\kappa > 0$. A random variable T is said to follow log-logistic distribution, if its logarithm, follows the logistic distribution; that is T follows log-logistic distribution if $Y = \log(T)$ is a logistic distribution where T is time. The probability density function of a log-logistic is given by

$$f(t) = \frac{\lambda\kappa(\lambda t)^{\kappa-1}}{(1 + (\lambda t)^\kappa)^2}, \quad t > 0.$$

The CDF for the log-logistic distribution is given by

$$F(t) = \frac{(\lambda t)^\kappa}{1 + (\lambda t)^\kappa}.$$

From the relation in Equation (3.6.9), the survival function is,

$$S(t) = \frac{1}{1 + (\lambda t)^\kappa}. \quad (3.6.22)$$

Hence the hazard function is given by

$$h(t) = \frac{\lambda\kappa(\lambda t)^{\kappa-1}}{1 + (\lambda t)^\kappa}. \quad (3.6.23)$$

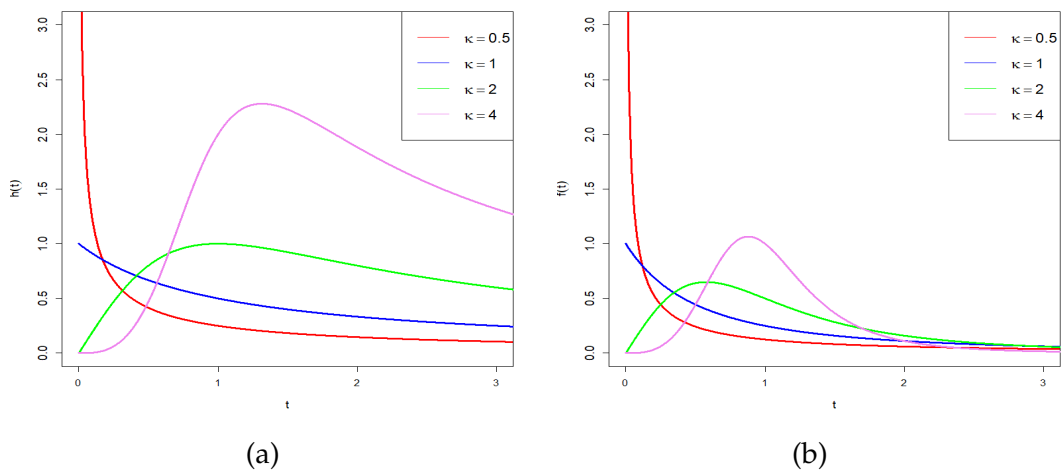


Figure 3.12: Figure 3.12a shows a plot of the hazard function $h(t)$ and Figure 3.12b shows a probability density function $f(t)$ for the log-logistic distribution for different κ 's and $\lambda = 1$.

From Figure 3.12a, it can be seen that the hazard of the log-logistic distribution starts from 0 and reaches its maximum at time $t = \frac{(\kappa-1)^{\frac{1}{\kappa}}}{\lambda}$ when $\kappa > 1$, which later declines as t approaches infinity. Also, the hazard of the log-logistic starts at λ but only decreases as t approaches infinity when $\kappa = 1$. Finally, the hazard starts from infinity and decreases when $\kappa < 1$. This distribution can be used to describe survival problems that have its hazard increasing at the initial stage and reducing at the later stage (Lee and Wang, 2003).

3.6.3.4 Log-normal distribution

The log-normal distribution of a continuous random variable T is characterised by two parameters. This distribution is most easily characterized by saying the continuous random variable T follows log-normal distribution if the logarithm of T is normally distributed; that is, T follows log-normal distribution if $Y = \log(T)$ is normally distributed with mean and variance specified by μ and σ^2 respectively. Hence, Y is of the form $Y = \mu + \sigma Z$, where Z is a standard normal. It is therefore expressed by $T \sim \text{Log-normal}(\mu, \sigma^2)$. Its probability density function is hence given by

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln t - \mu)^2}{2\sigma^2}\right), \quad t > 0.$$

Its cumulative density function is also given by

$$F(t) = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{\sqrt{2}(\ln t - \mu)}{2\sigma}\right),$$

where

$$\operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t \exp(-u^2) du.$$

Using the relation in Equation (3.6.9) the survival function can be written as

$$S(t) = \frac{1}{2} - \frac{1}{2} \operatorname{erf}\left(\frac{\sqrt{2}(\ln t - \mu)}{2\sigma}\right).$$

Therefore, the hazard function is expressed as

$$h(t) = \frac{\sqrt{2} \exp\left(-\frac{(\ln t - \mu)^2}{2\sigma^2}\right)}{t\sigma\sqrt{\pi} \left(1 - \operatorname{erf}\left(\frac{\sqrt{2}(\ln t - \mu)}{2\sigma}\right)\right)}. \quad (3.6.24)$$

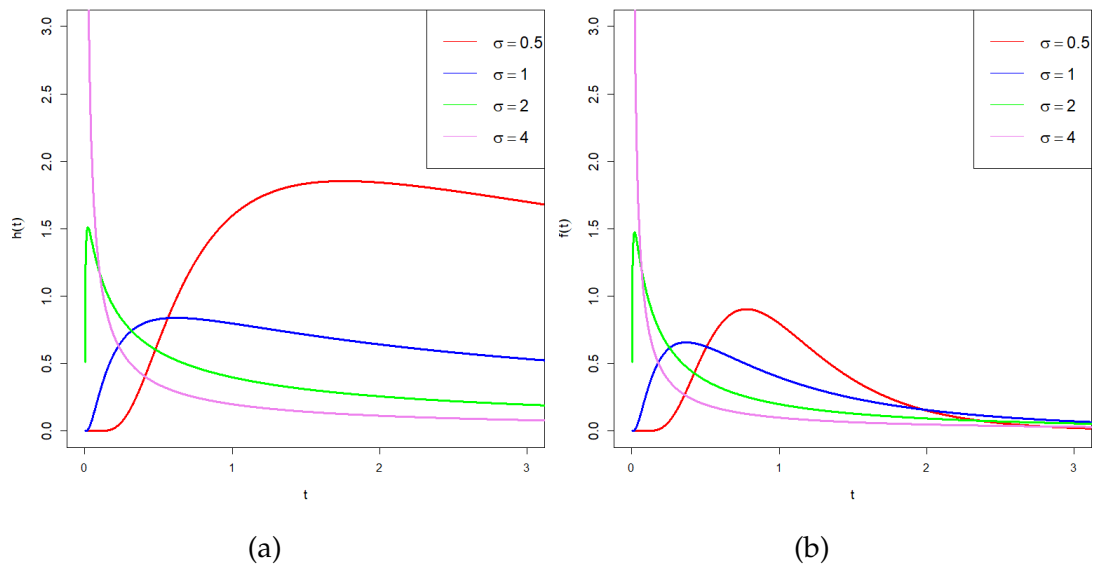


Figure 3.13: Figure 3.13a shows a plot of the hazard function $h(t)$ and Figure 3.13b shows a probability density function $f(t)$ for the log-normal distribution for different σ 's and $\mu = 1$.

Lee and Wang (2003) stated that log-normal distribution best describes survival problems with initial increase of the hazard rate to a peak and later decrease of the hazard rate to zero as t approaches infinity.

3.6.3.5 Extreme-value distribution

The extreme-value distribution of a continuous random variable T is characterised by two parameters. They are the location parameter λ and scale parameter $\delta > 0$. The probability density function of the extreme-value distribution is expressed as

$$f(t) = \frac{1}{\delta} \exp \left[\frac{t - \lambda}{\delta} - \exp \left(\frac{t - \lambda}{\delta} \right) \right], \quad -\infty < t < \infty.$$

The CDF for the extreme-value distribution is given by

$$F(t) = 1 - \exp \left[- \exp \left(\frac{t - \lambda}{\delta} \right) \right].$$

therefore we can deduce the survival function as

$$S(t) = \exp \left[- \exp \left(\frac{t - \lambda}{\delta} \right) \right].$$

Hence computing the hazard function, we get,

$$h(t) = \frac{1}{\delta} \exp\left(\frac{t - \lambda}{\delta}\right). \quad (3.6.25)$$

Survival parametric regression models

Linear regression seeks to identify a relationship that exists in the dependent variable (y_i) and independent variable (x_i). However, linear regression is able to achieve this by satisfying certain underlying assumptions. This study is also interested in investigating whether there exists some relationship between survival time and explanatory variables. For instance, suppose t denotes survival time and $\mathbf{x} = (x_1, \dots, x_m)^T$ a vector of differentially expressed genes, then we investigate if the amount of genes expressed has an effect on the survival time. For this study, these explanatory variables can also be referred to as prognostic factors. Survival regression models are mostly expressed in terms of hazard function and are given by

$$\begin{aligned} h(t, \mathbf{x}) &= \exp(\beta_0 + \beta \mathbf{x}), \\ h(t, \mathbf{x}) &= \exp(\beta_0) \exp(\beta \mathbf{x}). \end{aligned} \quad (3.6.26)$$

where \mathbf{x} are explanatory variables, $\mathbf{x} \in \mathbb{R}^d$, $\beta = [\beta_1, \beta_2, \dots, \beta_d]$ are also coefficient of the explanatory variables, $\beta \in \mathbb{R}^d$ and β_0 is a baseline hazard function. Denoting $h_0(t) = \exp(\beta_0)$, Equation (3.6.26) becomes

$$h(t, \mathbf{x}) = h_0(t) \exp(\beta \mathbf{x}). \quad (3.6.27)$$

$h_0(t)$ is called the baseline hazard function because, $h(t, \mathbf{x})$ reduces to $h_0(t)$ when there are no explanatory variables; that is when $\mathbf{x} = 0$ (Kleinbaum, 1998).

From Equation (3.6.19) and Equation (3.6.27), we can also write a survival regression model as

$$S(t, \mathbf{x}) = \exp\left[\left(-\int_0^t h_0(u) du\right) \exp(\beta \mathbf{x})\right]. \quad (3.6.28)$$

Equation (3.6.28) can be expanded to give a new expression written as

$$S(t, \mathbf{x}) = \left[S_0(t)\right]^{\exp(\beta \mathbf{x})}, \quad (3.6.29)$$

where $S_0(t) = \exp\left(-\int_0^t h_0(u)du\right)$ is called the baseline survival function.

Given that the probability distribution function is known, we can then substitute the $h_0(t)$ with the hazard function of the corresponding probability distribution. Using the Weibull distribution as an example, the hazard function is known from Equation (3.6.21). Hence, the hazard function for the Weibull regression model can be written using Equation (3.6.27) as

$$h(t, \mathbf{x}) = \frac{\gamma}{\alpha} t^{\gamma-1} \exp(\beta \mathbf{x}). \quad (3.6.30)$$

Similarly, the survival function for the Weibull survival regression model can be expressed by using Equation (3.6.29) as

$$S(t, \mathbf{x}) = \left[\exp\left(-\frac{1}{\alpha} t^\gamma\right) \right]^{\exp(\beta \mathbf{x})}. \quad (3.6.31)$$

Also, while linear regression uses the coefficient of the explanatory variable β as its measure of effect, survival analysis uses **Hazard Ratio (HR)** as its measure of effect. The hazard ratio is simply an expression written in terms of an exponential of one or more coefficients of the explanatory variable of a model (Kleinbaum, 1998). It is actually computed by dividing the hazard of an individual by the hazard of another individual.

Considering two individuals with explanatory variables \mathbf{x}_1 and \mathbf{x}_2 respectively, then HR is given by

$$\begin{aligned} \text{HR} &= \frac{h_1(t, \mathbf{x}_1)}{h_2(t, \mathbf{x}_2)} = \frac{h_0(t) \exp(\beta \mathbf{x}_1)}{h_0(t) \exp(\beta \mathbf{x}_2)}, \\ \text{HR} &= \exp[\beta(\mathbf{x}_1 - \mathbf{x}_2)]. \end{aligned} \quad (3.6.32)$$

$\text{HR} = 1$ means that both individuals have equivalent hazards. Also, $\text{HR} > 1$ implies that the first individual has a higher risk than the second individual. Similarly, $\text{HR} < 1$ implies that the first individual has a lower risk than the second individual.

The only limitation for the survival parametric regression model is, the underlying probability distribution is mostly unknown hence making it difficult to obtain the exact form of the model (Lee and Wang, 2003). Zhang (2016) stated that the underlying mathematical model of the survival time needs to be described

but to a great extent may be unrealistic and stringent. Fortunately, there are techniques that can help determine or approximate the model but do not require the underlying probability distribution to be known.

3.6.4 Semi-parametric survival models

A semi-parametric model does not specify the underlying probability distribution of the survival time but also possess a parametric property. A popular example is the Cox proportional hazard which takes the form of Equation (3.6.27).

Describing Cox proportional hazard model

Generally, the Cox proportional hazard model with explanatory variables $\mathbf{x} = (x_1, \dots, x_m)^T$ is of the form,

$$h(t, \mathbf{x}) = h_0(t) \exp(\beta \mathbf{x}). \quad (3.6.33)$$

Cox proportional hazard do not specify the probability distribution of the baseline hazard ($h_0(t)$). For this reason, Cox proportional hazard is widely used if the underlying assumptions of the baseline hazard are not of interest. Also, the results of Cox proportional hazard are very close to correctly approximated parametric models hence making them very robust. Cox proportional hazard can also be used to estimate the hazard ratio discuss above. Specifically, Cox proportional hazard can be used to compare and evaluate the relative risks of patients. This is done by dividing the risk of a patient by a baseline risk. The baseline risk is evaluated by multiplying thee coefficients of the variables with the average of the genes.

3.6.5 Statistical model selection

A study like this will need a statistical model from the explanatory variables upon which the hazard function will depend. Statistical models are used to depict the distribution underlying the data used. Such models provide useful information about how the data was generated, but they are almost never exact. This is because some information may be lost by using statistical models to depict the data generation process.

It is therefore important to know the significant explanatory variables that improve the prediction power of the hazard function. By knowing the significant explanatory variables, several alternative models can be built and the best among them should be chosen. For example, two models A and B may have explanatory variables that are all contributing to the predictive power of the hazard function. In such a case, the model with the highest predictive power is preferred. There can also be cases where the explanatory variables of model A is a subset for that of model B . In this case, we say model A is nested in model B and the two models are said to be parametrically nested (Kleinbaum, 1998).

For these reasons, the study employs statistical model selection methods to help compare and fit the best model among alternatives. In survival analysis, the two most common statistics used to compare alternative models are the log-likelihood ratio and Akaike Information Criterion (AIC) (Collett, 1993).

The log-likelihood ratio compares whether there is a significant difference between parametric nested models. Thus, it is used to compare the significance of an additional variable between the two models. This statistic employs a likelihood function which summarises the information of a model as $-2 \ln(L(\beta))$. Hence, the statistic is given by

$$-2[(\ln L(\beta)_{\text{modelA}}) - (\ln L(\beta)_{\text{modelB}})], \quad (3.6.34)$$

where β is a vector of coefficients of the variables in the model and $L(\beta)$ is an estimate of the maximum likelihood function given by

$$L(\beta) = \prod_{j=1}^r \frac{\exp(\beta^T x_j)}{\sum_{t \in R(t_j)} \exp(\beta^T x_t)}, \quad (3.6.35)$$

where β^T is a transpose of the vector of coefficients of the variables.

Furthermore, it asymptotically follows a chi-square distribution with a null hypothesis that β for additional explanatory variables equals to zero. The number of degrees of freedom for this distribution is given by the difference between the number of independent β -parameters being fitted under the two models. Two models are therefore considered equivalently significant if their log-likelihood ratio is significantly small, but for simplicity, the model with a smaller number of explanatory variables is preferred. In the case where the log-likelihood ratio

for two models is significantly large, the additional explanatory variables are required to improve the predictive power of the hazard function (Collett, 1993).

The AIC is also used to estimate the relative quality of statistical models by finding the information lost in a model. AIC is therefore given by

$$\text{AIC} = 2k - 2\ln(L(\hat{\beta})), \quad (3.6.36)$$

where k is the number of parameters to be estimated and $L(\hat{\beta})$ is the estimate of the maximum likelihood function given in Equation (3.6.35). Models do not have to be necessarily nested to apply AIC (Kleinbaum, 1998). The best model is then chosen based on the model with the smallest AIC value. For this study, AIC will be used to find the best probability distribution for the survival time while the log-likelihood ratio will be used to find the best model in a parametrically nested model.

3.7 Software and packages used

The software used for this research is the R statistical program. The packages used in the software also includes RTCGA, samr, randomForest, keras, e1071, lhs, survival, and survminer. RTCGA was used to obtain TCGA gene expression and clinical data. samr was also used to find genes that are differentially expressed. randomForest package was used to analysis random forests. keras and e1071 were used to analyse ANN and SVM respectively. Latin hypercube samples were also obtained by using the package lhs. survival and survminer were also used for survival analysis and fancy plotting respectively.

In summary, this chapter highlighted the techniques to be used for the study. It started with the type of data (RNA-seq) to be used for the study and where it was obtained from. It further presented an algorithm (SAM) which is used to identify genes that are differentially expressed. Three machine learning algorithms were also presented which will be applied to classify breast cancer patients. Finally, survival analysis was described which will also be used to build a model to predict how long a breast cancer patient survives based on their differentially expressed genes.

Chapter 4

Statistical analysis of breast cancer gene expression and clinical data

This chapter seeks to analyse breast cancer gene expression and clinical data. It first identifies genes that are differentially expressed by applying Significance Analysis of Microarray (SAM) to the TCGA data. The differentially expressed genes identified are then used to predict breast cancer patients using three machine learning methods. Finally, a model is built to predict survival time for breast cancer patients.

4.1 Identifying essential genes

The study seeks to identify genes that contribute to the formation of breast cancer. The algorithm used to achieve this is called the Significant Analysis of Microarray (SAM). For a set of genes, SAM identifies genes that are differentially expressed by defining a threshold Δ for which genes with scores greater than Δ are deemed potentially significant. The algorithm also returns a quantile–quantile (Q–Q) plot between the observed scores of genes and the expected scores of genes. As a reminder, the observed score is obtained from using the whole data while the expected score is also obtained from permuting the response values.

The algorithm, therefore, re–sampled the data 19 times to obtain observed scores for the genes. It also performed 100 permutations on the response variable to obtain the expected score for each gene. To be specific, observed scores became

Chapter 4. Statistical analysis of breast cancer gene expression and clinical data

52

stable after re-sampling 19 times. Also, expected scores were invariant after 100 permutations. This yielded a maximum $\Delta = 49986$ with 23 positive genes called significant and zero negative genes. The results can be visualised in Figure 4.1.

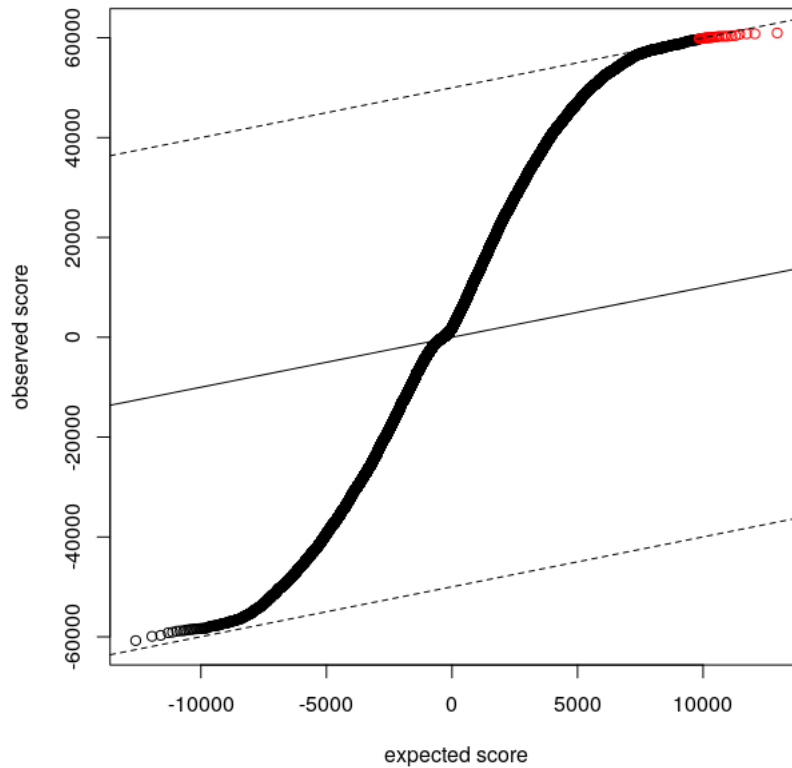


Figure 4.1: The Q-Q plot obtained by using the SAM algorithm with $\Delta = 49986$. Scores obtained by re-sampling 19 times and permuting 100 times. Red points represent differentially expressed genes and black points represent non-differentially expressed genes

The black points in Figure 4.1 represent genes that have no difference in their observed scores and their expected scores. For this reason, they are deemed not significant. The red points represent genes with observed scores significantly different from their expected scores hence they are called significant genes. Specifically, they are significantly positive genes. The results did not contain any negatively significant gene. The significant genes in descending order of rank score include *FIGF*, *SDPR*, *CD300LG*, *ADAMTS5*, *MAMDC2*, *TMEM220*, *SPRY2*, *PAMR1*, *ARHGAP20*, *TSLP*, *LMOD1*, *FAM13A*, *CA4*, *PPP1R12B*, *ABCA10*, *MME*, *SCN4B*, *DMD*, *FXVD1*, *CAV1*, *ITIH5*, and *BTNL9*. Table 4.1, therefore, shows significant genes with their corresponding description and scores. In addition,

Table 4.2 also shows details of the molecular function of each gene and its corresponding biological process. These details were obtained from GeneCards® (<https://www.genecards.org/>) and UniProt (<https://www.uniprot.org/>).

Table 4.1: Significant genes with their corresponding description and scores

Gene Symbol	Description	Score(T)
<i>FIGF</i>	Vascular Endothelial Growth Factor D	60960.833
<i>SDPR</i>	Serum Deprivation Response	60807.111
<i>CD300LG</i>	Cluster of Differentiation 300 Molecule Like Family Member G	60801.778
<i>ADAMTS5</i>	A Disintegrin And Metalloproteinase With Thrombospondin Motifs-5	60667.889
<i>MAMDC2</i>	MAM(Meprin, A-5 protein, and receptor protein-tyrosine phosphatase Mu) Domain Containing 2	60440.667
<i>TMEM220</i>	Transmembrane Protein 220	60436.278
<i>SPRY2</i>	Sprouty RTK Signaling Antagonist 2	60308.778
<i>PAMR1</i>	Peptidase Domain Containing Associated With Muscle Regeneration 1	60308.389
<i>SCARA5</i>	Scavenger Receptor Class A Member 5	60295.444
<i>ARHGAP20</i>	Rho GTPase Activating Protein 20	60218.778
<i>TSLP</i>	Thymic stromal lymphopoietin	60144.842
<i>LMOD1</i>	Leiomodin 1	60175.737
<i>FAM13A</i>	Family With Sequence Similarity 13 Member A	60161.737
<i>CA4</i>	Carbonic Anhydrase 4	60067.947
<i>PPP1R12B</i>	Protein Phosphatase 1, Regulatory (Inhibitor) Subunit 12B	60021.611
<i>ABCA10</i>	ATP Binding Cassette Subfamily A Member 10	60004.158
<i>MME</i>	Membrane Metalloendopeptidase	59994.333
<i>SCN4B</i>	Sodium Voltage-Gated Channel Beta Subunit 4	59983.222
<i>DMD</i>	Dystrophin	59893.778
<i>FXYP1</i>	FXYP Domain Containing Ion Transport Regulator 1	59884.833
<i>CAV1</i>	Caveolin 1	59881.056
<i>ITIH5</i>	Inter-Alpha-Trypsin Inhibitor Heavy Chain Family Member 5	59852.389
<i>BTNL9</i>	Butyrophilin Like 9	59832.667

Chapter 4. Statistical analysis of breast cancer gene expression and clinical data 54

Table 4.2: Molecular function and biological process for significant genes

Gene Symbol	Molecular Function	Biological Process
<i>FIGF</i>	Growth factor activity	Angiogenesis, cell proliferation, positive regulation of cell division
<i>SDPR</i>	Protein kinase C binding	Plasma membrane tubulation
<i>CD300LG</i>	Receptor	Regulation of immune response
<i>ADAMTS5</i>	Extracellular matrix binding	Defense response to bacterium
<i>MAMDC2</i>	–	–
<i>TMEM220</i>	–	–
<i>SPRY2</i>	Protein serine/threonine kinase activator activity	Regulation of cell differentiation, negative regulation of angiogenesis
<i>PAMR1</i>	Calcium ion binding	May play a role in regeneration of skeletal muscle
<i>SCARA5</i>	Scavenger receptor activity	Cellular iron ion homeostasis
<i>ARHGAP20</i>	GTPase activation	Regulation of small GTPase mediated signal transduction
<i>TSLP</i>	cytokine activity, interleukin-7 receptor binding	Positive regulation of inflammatory response, negative regulation of apoptotic process
<i>LMOD1</i>	Actin-binding	Muscle contraction, positive regulation of actin filament polymerization
<i>FAM13A</i>	GTPase activation	Regulation of small GTPase mediated signal transduction
<i>CA4</i>	Carbonate dehydratase activity, zinc ion binding	Bicarbonate transport
<i>PPP1R12B</i>	protein kinase binding	signal transduction
<i>ABCA10</i>	ATP binding	lipid transport
<i>MME</i>	endopeptidase activity	cellular response to cytokine stimulus
<i>SCN4B</i>	voltage-gated sodium channel activity	regulation of sodium ion transmembrane transporter activity
<i>DMD</i>	actin binding	regulation of skeletal muscle contraction
<i>FXYP1</i>	sodium channel regulator activity	sodium ion transport
<i>CAV1</i>	ATPase binding	apoptotic signaling pathway
<i>ITIH5</i>	serine-type endopeptidase inhibitor activity	hyaluronan metabolic process
<i>BTNL9</i>	–	–

“–” means gene do not have either known biological process or known molecular function

4.2 Classifying patients based on essential genes

Using labelled response variable from the data, the machine learning algorithms, including random forests, artificial neural network (ANN) and support vector machines (SVM) were able to classify breast cancer patients into two classes

namely; a tumour or a tumour-free class. The classifications were based on the significant genes identified by SAM. The data matrix used had a dimension of 1212×24 . 1212 represented the number of patients while 24 represented 23 significant genes and one response variable.

The data matrix was then randomly divided into training data and test data. 70% of the data were allocated to the training data while the remaining 30% were assigned to the test data. To obtain optimal results, the trained models had to be tuned. They were tuned using a validation data which was generated from the training data. Specifically, the random forest was tuned using the OOB samples, whereas ANN and SVM used 20% of the training data to validate the trained model.

After several tuning, random forests used 200 trees with 4 splitting variables at each node to build a model for the study. Similarly, the artificial neural network was built with 5 layers. It had an input layer, an output layer and 3 hidden layers. The input layer had 128 units, while the output layer had only two units. However, the first hidden layer had 32 units followed by 16 units in the second hidden layer and the last hidden layer had 8 units. In addition, each layer had a bias except for the output. Thus, there were 7882 units in total for the artificial neural network. The activation function used for each layer was ReLU, except for the output layer which used the softmax activation function. Also, for the support vector machine, the radial basis was better than the other three kernel functions discussed.

Their performances were measured using a confusion matrix. Again, confusion matrix measures how well an algorithm categorised each patient into a positive group or negative group. The positive group for this study is the tumour class, whereas the negative is a tumour-free class. This procedure was also applied for the median supplement explained in Section 3.4.4.

4.2.1 Comparing the performance of classification algorithms with and without supplement data

First, Table 4.3 summarises the performance of the machine learning algorithms without the supplement data (that is unbalanced data with 1100 labelled tumour and 112 labelled tumour-free).

Chapter 4. Statistical analysis of breast cancer gene expression and clinical data 56

Table 4.3: Comparison of the performance of the classifiers without supplement data

	Sensitivity	Specificity	False Positive	False Negative	Accuracy
Random Forest	0.99377	0.96552*	0.03448*	0.00623	0.99143*
Neural Network	0.97708	0.91176	0.08824	0.02292	0.97128
Support Vector	0.99685*	0.85714	0.14286	0.00315*	0.98551

“*” means preferred algorithm for the statistical measure

The results from Table 4.3 indicates that random forests was the best of the three because it had the highest rate in accuracy. The next classifier with a higher rate of accuracy was SVM and finally followed by the neural network. The performance of the neural network may be due to the fact that the data is unbalanced. Using random forest without the supplement data, we can say that, 99.377% of the patients may be correctly diagnosed with breast cancer. Also, 96.552% of the patients may be correctly classified as not having breast cancer. 3.448% of the patients may be falsely diagnosed with breast cancer. Finally, 0.623% of the patients may be falsely classified as not having breast cancer.

Similarly, Table 4.4 also summarises performance of the machine learning algorithms after using median supplement approach. After using the median supplement technique to balance the data, the dimension of the data matrix changed to 2200×24 .

Table 4.4: Comparison of the performance of the classifiers using median supplement data

	Sensitivity	Specificity	False Positive	False Negative	Accuracy
Random Forest	0.94817*	1.00000*	0.00000*	0.05183*	0.97331*
Neural Network	0.84404	0.84953	0.15047	0.15596	0.84675
Support Vector	0.80060	0.97444	0.02556	0.19940	0.88509

“*” means preferred algorithm for the statistical measure

Again, the results from Table 4.4 indicates that random forests performed better than the other classifiers because it had the highest rate of accuracy. The next highest rate of accuracy was the SVM and finally followed by the neural network. Again, the results from the random forests with median supplement data indicate that 94.817% of the patients may be truly diagnosed with breast cancer. Interestingly, 100% of the patients may be correctly classified as not having breast cancer. Also, no patient may be wrongly diagnosed with the disease. Finally, 5.183% of the patients may be falsely classified as not having breast cancer.

In addition, a mean supplement was also generated to investigate if there would

be a difference in the results. Table 4.5 summarises performance of the machine learning algorithms after using the mean supplement approach. Similarly, the dimension of the data matrix changed to 2200×24 after using the mean supplement technique.

Table 4.5: Comparison of the performance of the classifiers using mean supplement data

	Sensitivity	Specificity	False Positive	False Negative	Accuracy
Random Forest	0.95918*	1.00000*	0.00000*	0.04082*	0.97901*
Neural Network	0.88889	0.94328	0.05672	0.11111	0.97264
Support Vector	0.90801	0.98452	0.01548	0.09199	0.94545

“*” means preferred algorithm for the statistical measure

Once again, random forests had higher rates in sensitivity, specificity and accuracy, hence it performed better than the other classifiers. Using random forest with mean supplement data we can say that, 95.918% of the patients may be truly diagnosed with breast cancer. Also, any patient diagnosed as not having breast cancer may be true. Again, no patient may be wrongly diagnosed with breast cancer. Finally, 4.082% of the patients may be wrongly classified as not having breast cancer.

4.2.2 Details of the best performing classification algorithm

By far, the random forest has proven to be the best classifier of the data, hence a deeper exploration will be beneficial. First, it is relevant to know the variables that increase the predictive power of the model. Identifying these relevant genes can help fine-tune the model to increase the prediction power. This is achieved plotting the average decrease of Gini index that each gene contributes to the model. Figure 4.2 summarises the first 10 important variables.

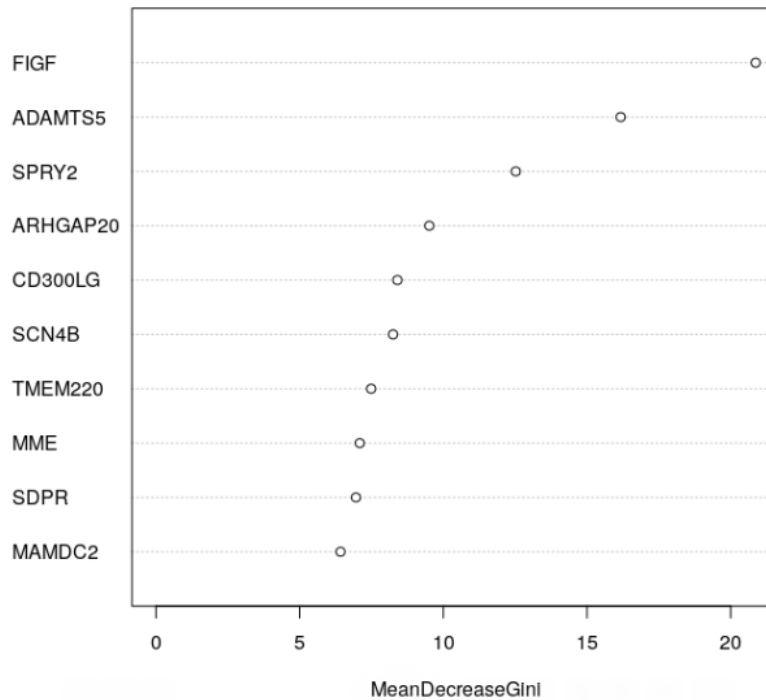


Figure 4.2: First 10 variables that increase the prediction power of the random forest

Figure 4.2 shows that *FIGF* has the most averaged Gini index decrease of error for the model. Interestingly, in the SAM analysis, *FIGF* also had the highest score. The first 10 important variables were then used to build a new model. Fortunately, there was no significant improvement of results from the new model. Occam's razor states that "among several plausible explanations for a phenomenon, the simplest is best" (Faraway, 2002). Thus, the final model to predict breast cancer will be based on the model with fewest variables which is the first 10 most important variables. The new model presents a tree displayed in Figure 4.3.

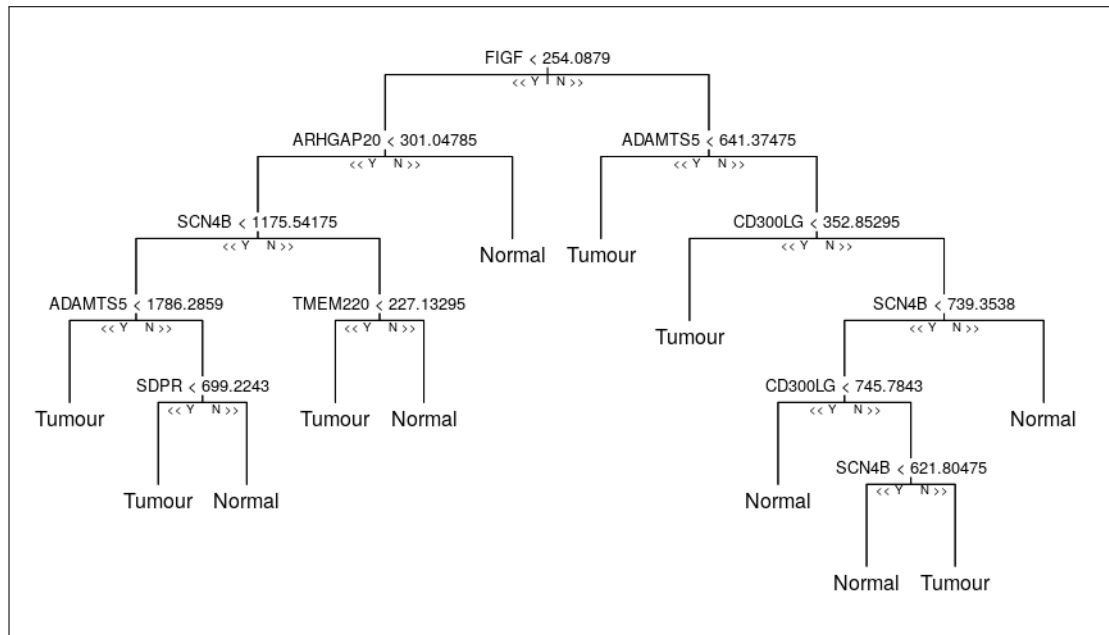


Figure 4.3: A random tree obtained for random forest algorithm with 23 nodes. 12 of the nodes are leaf nodes and 10 of them are internal nodes. Y means **yes** and proceed to the left while N means **no** and proceed to the right.

Figure 4.3 shows a random tree constructed by the model. This tree is relatively small because it had a root node, 10 internal nodes and 12 leaf nodes. With this tree, breast cancer can easily be predicted or diagnosed using a patient's gene expression data.

4.3 Identifying association between essential genes and patients survival

Next is to build a model that predicts the survival time of patients using their clinical data and the 23 significant genes. The clinical data contains both patients who experience the event of interest (death) and patients who were lost to follow-up or were still alive at the end of the study (censored). Among the patients, 104 had experienced the event of interest while 993 were censored. However, the survival time for breast cancer patients in this study ranges from 1 to 7067. Figure 4.4 shows only the survival time of the first 50 patients for the study.

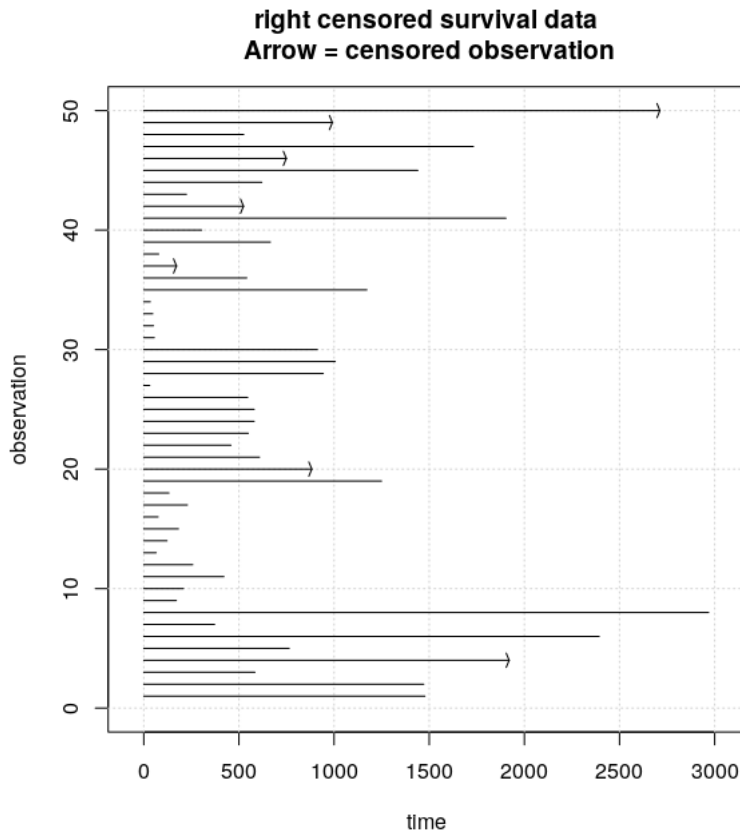


Figure 4.4: Plot showing the first 50 patients and their survival time. Arrowed bars indicates a censored patient.

4.3.1 Kaplan–Meier survival analysis

Figure 4.4 does not convey much information as it only reveals the survival time of some patients. However, the Kaplan–Meier curve is able to estimate the survival probability of patients given their survival time. Figure 4.5 displays a Kaplan–Meier curve for the patients.

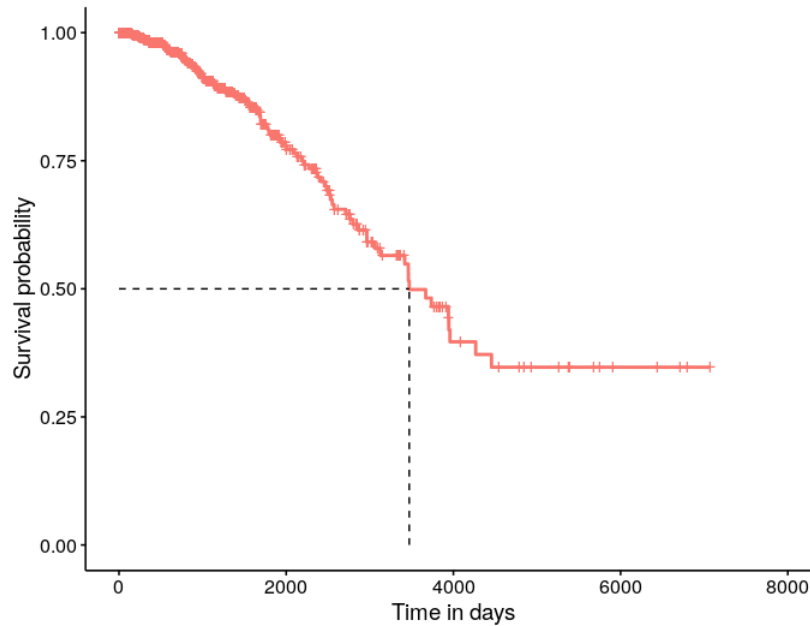


Figure 4.5: A Kaplan–Meier curve to estimate the survival probability of the patients with a median survival time at 3472 days. The vertical lines crossing the curves represent censored patients.

From Figure 4.5, it can be seen that the survival probability reduces as time increases. At the time $t = 0$, the survival probability was 100% because at $t = 0$, no patient had experienced the event of interest. The Kaplan–Meier then estimated a median survival time of patients to be 3472 days (that is approximately 9 years and 6 months). This means that the probability that a breast cancer patient will survive beyond 3472 days is 50%.

Furthermore, the study investigated the influence of variables (significant genes) on the survival probability. It tested whether the level (high or low) of gene expression could affect the survival probability. For example, if a patient expresses a high level of *FIGF*, what is the probability that she will survive beyond a particular time. For this reason, the study employs the log–rank test statistic to investigate the influence gene levels expressed has on survival probability.

Table 4.6: Log-rank test statistic for genes

Gene Symbol	Chi-square	<i>p</i> -value
<i>FIGF</i>	2.77	0.0960
<i>SDPR</i>	9.38	0.0022*
<i>CD300LG</i>	3.78	0.0519
<i>ADAMTS5</i>	3.97	0.0464*
<i>MAMDC2</i>	2.50	0.1138
<i>TMEM220</i>	4.90	0.0268*
<i>SPRY2</i>	4.40	0.0360*
<i>PAMR1</i>	8.10	0.0044*
<i>SCARA5</i>	1.48	0.2245
<i>ARHGAP20</i>	1.77	0.1831
<i>TSLP</i>	8.67	0.0032*
<i>LMOD1</i>	2.82	0.0930
<i>FAM13A</i>	0.70	0.4036
<i>CA4</i>	0.64	0.4255
<i>PPP1R12B</i>	0.06	0.8012
<i>ABCA10</i>	2.26	0.1328
<i>MME</i>	3.02	0.0821
<i>SCN4B</i>	1.40	0.2375
<i>DMD</i>	2.27	0.1321
<i>FXVD1</i>	1.02	0.3132
<i>CAV1</i>	5.07	0.0243*
<i>ITIH5</i>	1.41	0.2343
<i>BTNL9</i>	8.48	0.0036*

*" means statistically significant

Table 4.6 thus summarises the influence of variables on patients survival probability. The table presents genes with their corresponding chi-square (χ^2) and *p*-values. Using χ^2 with one degree of freedom, the *p*-values for the genes were generated. Results indicate that *SDPR*, *ADAMTS5*, *TMEM220*, *SPRY2*, *PAMR1*, *TSLP*, *CAV1*, and *BTNL9* are significantly different. This means that their levels of expression affect patients survival probability. In details, Kaplan-Meier curves have been plotted to display significant genes identified by the log-rank statistic.

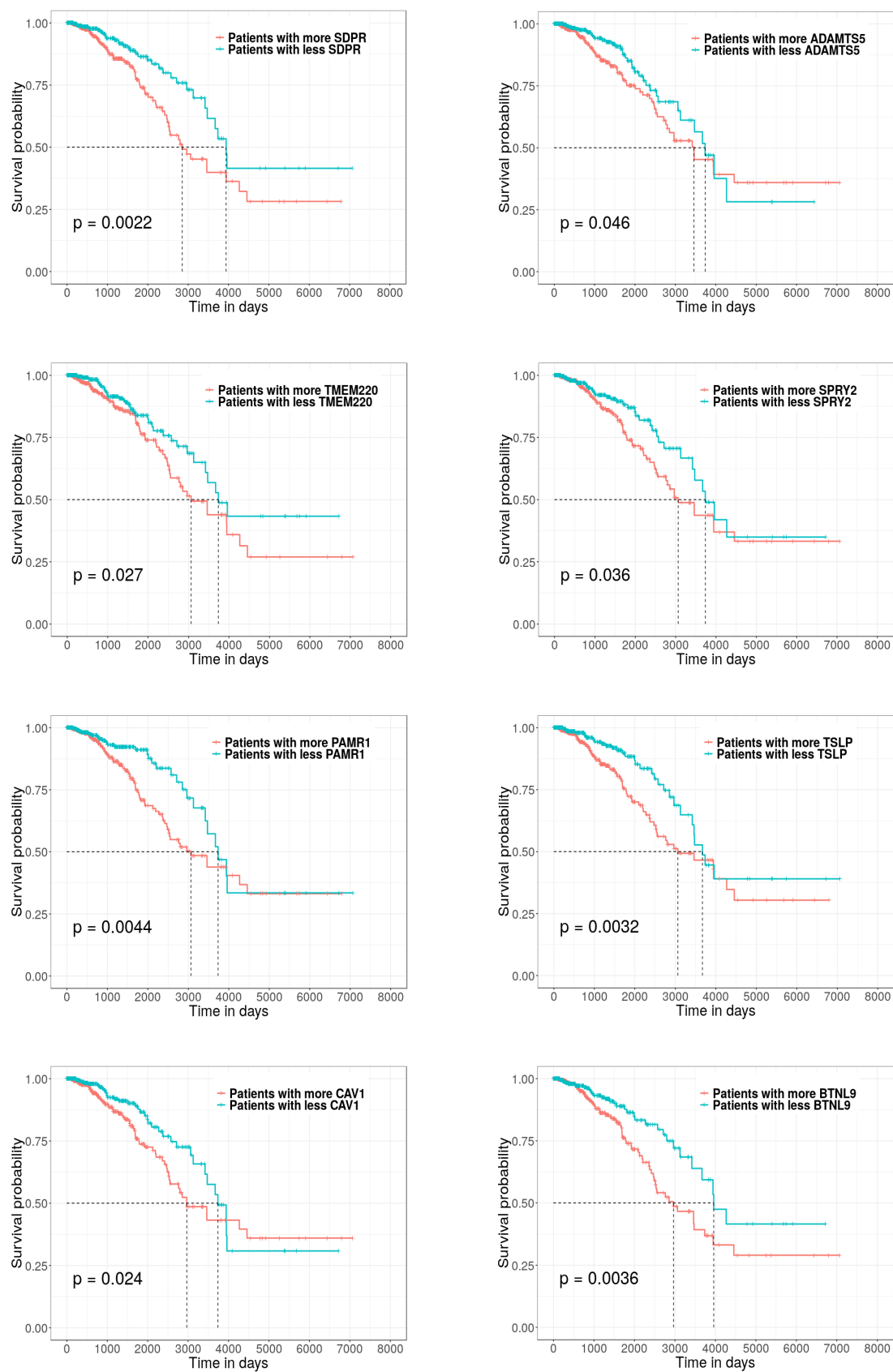


Figure 4.6: Plot for significant genes after using log-rank test statistics. Red curves indicate Kaplan-Meier curve for patients who express more and blue curves for patients who express less.

Figure 4.6 thus displays the Kaplan–Meier curve for each significant gene. For each plot of a significant gene, the red curve represents Kaplan–Meier curve for patients who express more of that particular significant gene. Similarly, the blue curve also represents the Kaplan–Meier curve for patients who express less of a particular significant gene. In general, it can be seen from the plots that the red curves were below the blue curves. Thus, it indicates that patients who expressed more of the significant genes had shorter survival time compared to patients who expressed less of the significant genes. This may be attributed to the fact that the significant genes identified by SAM were all significantly positive genes.

4.3.2 Building parametric survival models

As stated in Chapter 3, survival probability may depend on several factors. Unfortunately, Kaplan–Meier can estimate the survival probability with only one variable at a time. Parametric distributions can be used to build models to predict survival probability with more than one variable. Again, parametric distributions can be very accurate if it's underlying assumptions are met. To investigate parametric distributions, we first use AIC to identify the probability distribution that best describes the survival time.

Table 4.7 presents specific probability distributions used for the study with their corresponding AIC values. The results indicate that log–logistic distribution best describes the survival time for the patients. This is because it gave the least AIC value of 2050.11. Most importantly, this means that among the probability distributions log–logistic relatively provides the least information lost from the patient's survival time. It can also be seen that extreme–value distribution had the highest AIC value of 2198.34 therefore, it relatively lost more information than the other probability distributions. Figure 4.7 graphically presents the survival curves for the probability distributions for the study. It can be seen that log–logistic again is closer to the Kaplan–Meier curve than the other probability distributions. Even though Weibull was very close to the Kaplan–Meier curve but after about $t = 5000$ days, it deviated. It can thus be concluded that log–logistic distribution best describes the survival time of the patients.

Next, the study focuses on building a model for predicting the survival time

Table 4.7: Probability distributions with their corresponding AIC values

Probability Distribution	AIC values
Exponential	2094.11
Weibull	2053.34
Log-logistic	2050.11*
Log-normal	2055.25
Extreme	2198.34

*" means preferred probability distribution

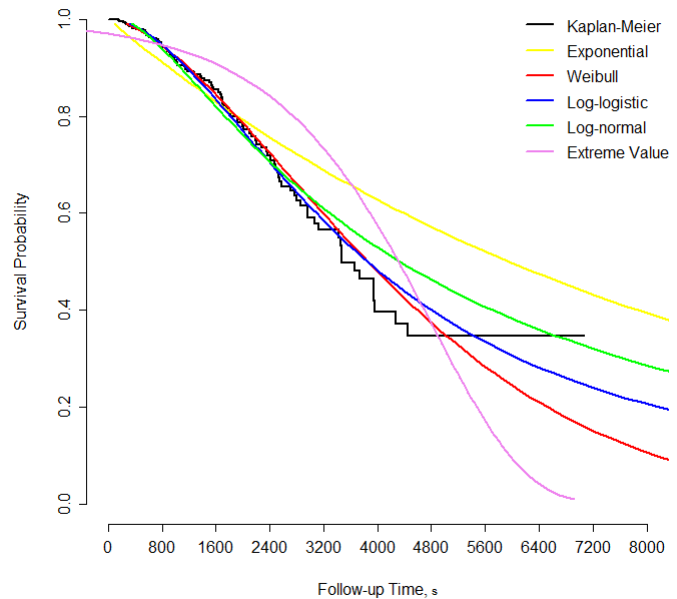


Figure 4.7: Figure 4.7 shows survival curves for the probability distributions.

for breast cancer patients. The model is built using a backward elimination approach to select variables that may be significant in the model. In addition, the statistical measure used for the variable selection is log-likelihood ratio. The variables used to build the model are the 23 significant genes identified by SAM. In essence, the model built after variable selection can be used to predict the survival probability of a patient. Table 4.8 provides details of the model with the 23 significant genes.

Table 4.8 provides details of building a full model with the 23 significant genes. Specifically, the table presents genes with their corresponding coefficients (Value), standard error (Std.Error), z-test value and its p -value. The p -value for the genes is generated from z-test which in turn is obtained by dividing the coefficient by its corresponding standard error. Thus, with 23 degrees of freedom and a significance level of 10%, the model yielded a p -value equal to 0.019 which resulted in a log-likelihood ratio of 39.19. The intercept is used to estimate the scale parameter (λ) while the Log(scale) is used to estimate the shape parameter (κ). Table 4.8 also shows that not all variables are significant. For example, only *FIGE*, *SDPR* and *PAMR1* will significantly contribute to prediction if the full

Table 4.8: Details of log–logistic model with the 23 significant genes

Gene Symbol	Value	Std.Error	z-test	p–value
(Intercept)	8.15e+00	1.60e-01	50.8483	0.00e+00
<i>FIGF</i>	-3.29e-03	1.70e-03	-1.9335	5.32e-02*
<i>SDPR</i>	-9.36e-04	4.17e-04	-2.2445	2.48e-02*
<i>CD300LG</i>	1.53e-03	1.57e-03	0.9763	3.29e-01
<i>ADAMTS5</i>	4.38e-04	4.10e-04	1.0696	2.85e-01
<i>MAMDC2</i>	-5.22e-04	4.97e-04	-1.0490	2.94e-01
<i>TMEM220</i>	-1.36e-03	1.62e-03	-0.8434	3.99e-01
<i>SPRY2</i>	2.64e-04	5.50e-04	0.4807	6.31e-01
<i>PAMR1</i>	-1.44e-03	4.72e-04	-3.0531	2.26e-03*
<i>SCARA5</i>	8.29e-04	9.20e-04	0.9014	3.67e-01
<i>ARHGAP20</i>	-5.08e-04	2.21e-03	-0.2296	8.18e-01
<i>TSLP</i>	2.95e-03	5.46e-03	0.5408	5.89e-01
<i>LMOD1</i>	2.21e-05	2.39e-04	0.0923	9.26e-01
<i>FAM13A</i>	3.15e-05	2.82e-04	0.1117	9.11e-01
<i>CA4</i>	-6.64e-04	3.29e-03	-0.2021	8.40e-01
<i>PPP1R12B</i>	7.64e-05	1.63e-04	0.4699	6.38e-01
<i>ABCA10</i>	-1.81e-03	2.69e-03	-0.6734	5.01e-01
<i>MME</i>	-2.06e-05	8.04e-05	-0.2562	7.98e-01
<i>SCN4B</i>	9.65e-04	6.52e-04	1.4799	1.39e-01
<i>DMD</i>	5.26e-04	3.30e-04	1.5904	1.12e-01
<i>FXVD1</i>	2.39e-03	1.93e-03	1.2386	2.15e-01
<i>CAV1</i>	-8.08e-05	1.03e-04	-0.7828	4.34e-01
<i>ITIH5</i>	2.01e-04	1.75e-04	1.1486	2.51e-01
<i>BTNL9</i>	-3.58e-04	8.05e-04	-0.4450	6.56e-01
Log(scale)	-6.77e-01	6.99e-02	-9.6793	3.69e-22

“*” means statistically significant

model is maintained because they are the only variables with p -value less than 10%. Thus a parametric nested model was built to improve the prediction of breast cancer patients.

Once again, the study employs a backward elimination procedure which removes variables with the largest p -value until all the variables have their p -value below the significance level. Table 4.9 gives details of how the variables were selected.

Table 4.9: Variable Selection Process for log-logistic distribution

Model	Log-likelihood ratio	<i>p</i> -value
model ₁ = Full model	39.19	1.9e-02
model ₂ = model ₁ - <i>LMOD1</i>	39.18	1.3e-02
model ₃ = model ₂ - <i>FAM13A</i>	39.17	9.4e-03
model ₄ = model ₃ - <i>CA4</i>	39.13	6.4e-03
model ₅ = model ₄ - <i>ARHGAP20</i>	39.08	4.3e-03
model ₆ = model ₅ - <i>MME</i>	39.03	2.8e-03
model ₇ = model ₆ - <i>BTNL9</i>	38.82	1.9e-03
model ₈ = model ₇ - <i>SPRY2</i>	38.53	1.3e-03
model ₉ = model ₈ - <i>PPP1R12B</i>	38.16	8.5e-04
model ₁₀ = model ₉ - <i>TSLP</i>	37.82	5.5e-04
model ₁₁ = model ₁₀ - <i>TMEM220</i>	37.49	3.5e-04
model ₁₂ = model ₁₁ - <i>MAMDC2</i>	37.07	2.2e-04
model ₁₃ = model ₁₂ - <i>SCARA5</i>	36.29	1.5e-04
model ₁₄ = model ₁₃ - <i>ABCA10</i>	35.68	9.6e-04
model ₁₅ = model ₁₄ - <i>CAV1</i>	34.49	7.3e-05
model ₁₆ = model ₁₅ - <i>ADAMTS5</i>	33.66	4.7e-05
model ₁₇ = model ₁₆ - <i>FXYD1</i>	32.70	3.0e-05
model ₁₈ = model ₁₇ - <i>CD300LG</i>	31.05	2.5e-05

Table 4.9 presents several models with their corresponding log-likelihood ratio and *p*-values. The log-likelihood ratio is used to generate the *p*-value. First, a full model containing all 23 significant genes was built. This resulted in a *p*-value of 0.019 indicating that the full model is significant under a 10% significance level. Next, the simplest method of the model was identified by backwards elimination. The process involved is to remove the gene with the largest *p*-value from the full model and fitting a new model without the removed gene. For example, *LMOD1* had the highest *p*-value of 0.926 (see Table 4.8) therefore, it was removed from the full model and new model was fitted without *LMOD1*. It can be seen that the model became more significant after *LMOD1* was removed. That is the global *p*-value of the full model reduced further from 0.019 to 0.013. This process of removing genes with the highest *p*-value and refitting continued until there was no significant improvement in the model. Thus, the last gene to be removed from the model was *CD300LG* with a *p*-value of 0.192. After *CD300LG* was removed from the model, there was no significant improvement in the model. The results of this is a reduced model presented in Table 4.10.

Table 4.10: Details of the reduced log–logistic model

Gene Symbol	Value	Std.Error	z-test	p-value
(Intercept)	8.127554	0.102452	79.33	0.00e+00
<i>FIGF</i>	-0.002298	0.001342	-1.71	8.68e-02*
<i>SDPR</i>	-0.000678	0.000163	-4.16	3.14e-05*
<i>PAMR1</i>	-0.001156	0.000377	-3.07	2.16e-03*
<i>SCN4B</i>	0.001195	0.000546	2.19	2.88e-02*
<i>DMD</i>	0.000491	0.000220	2.23	2.57e-02*
<i>ITIH5</i>	0.000229	0.000132	1.74	8.24e-02*
Log(scale)	-0.666871	0.069598	-9.58	9.55e-22

“*” means statistically significant

Table 4.10 presents genes that contribute significantly to predicting breast cancer patients survival probability. The table shows genes with their corresponding coefficient (Value), standard error (Std.Error), z–test value and p –value. It can be seen that all the genes now have their p –value less than the significant level which is 10%. The results indicate that *FIGF*, *SDPR*, *PAMR1*, *SCN4B*, *DMD* and *ITIH5* contribute significantly to predicting the survival probability of breast cancer patients. In addition, the reduced model resulted in an intercept of 8.127554 and Log(scale) of -0.666871. As discussed in Section 3.6.3, a random variable T follows log–logistic distribution if $Y = \log(T)$ is a logistic distribution. Thus, λ is estimated by $\exp(-(\text{Intercept}))$ therefore $\lambda = \exp(-8.127554) = 0.0003$. Also, κ is estimated by $(\exp(-\text{Log}(\text{scale})))^{-1} = (\exp(-0.666871))^{-1} = 1.9481$. Using Equation (3.6.23), the baseline hazard model at the time t is given as

$$h_0(t) = \frac{0.0003 \times 1.9481(0.0003 t)^{1.9481-1}}{1 + (0.0003 t)^{1.9481}}. \quad (4.3.1)$$

From Equation (4.3.1), the hazard model can now be written as

$$h(t, \mathbf{x}) = h_0(t) \exp(-0.002298 \text{ FIGF} - 0.000678 \text{ SDPR} - 0.001156 \text{ PAMR1} \\ + 0.001195 \text{ SCN4B} + 0.000491 \text{ DMD} + 0.000229 \text{ ITIH5}) \quad (4.3.2)$$

Equation (4.3.2) illustrates how an increase or decrease of a gene can affect the hazard model. For instance, an increase of expression of *FIGF* will decrease the hazard by a factor of $\exp(-0.002298)$; that is by 0.23%. Likewise, each increase of *SCN4B* will increase the hazard by a factor of $\exp(0.001195)$; that is by 0.12%.

Similarly, from Equation (3.6.22), the baseline survival function at time t is given

as

$$S_0(t) = \frac{1}{1 + (0.0003 t)^{1.9481}}. \quad (4.3.3)$$

Hence Equation (4.3.3) can be substituted into Equation (3.6.29) to obtain a survival model as

$$S(t, \mathbf{x}) = S_0(t) \exp(-0.002298 \text{ FIGF} - 0.000678 \text{ SDPR} - 0.001156 \text{ PAMR1} \\ + 0.001195 \text{ SCN4B} + 0.000491 \text{ DMD} + 0.000229 \text{ ITIH5}) \quad (4.3.4)$$

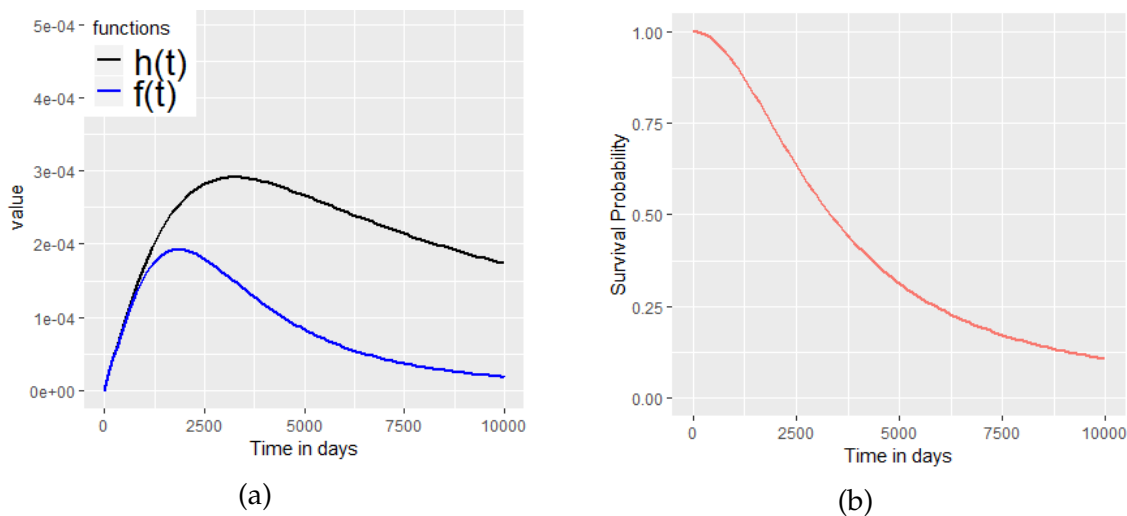


Figure 4.8: Using a scale parameter $\lambda = 0.0003$ and shape parameter $\kappa = 1.9481$, Figure 4.8a shows the hazard and probability density function for the log-logistic distribution. Figure 4.8b also shows the survival curve for the log-logistic distribution.

Figure 4.8 shows plots for hazard, probability density function and survival curve for log-logistic distribution. In Figure 4.8a, hazard initially increases but starts to decline after about time (t) equal to 3243.38 days. This was estimated by using $t = \frac{(\kappa-1)^{\frac{1}{\kappa}}}{\lambda}$ since $\kappa > 1$. This suggests that breast cancer patients with survival time less than 3243.38 days have a higher risk to the event than breast cancer patients with survival time more than 3243.38 days. In addition, the hazard suggests that breast cancer patients will experience the event of interest at a lower rate if they survive beyond $t > 3243.38$ days. Also in Figure 4.8b, the survival curve was very steep before $t = 3243.38$ but became gentle afterwards. This indicates a low survival probability for patients with survival time less than 3243.38 days.

4.3.3 Analysing the Cox proportional hazard model

Finally, the study uses Cox proportional hazard to explore the relative risk of the breast cancer patients. Similarly, a full model is built using the 23 significant genes identified by SAM. The backward elimination method of variable selection is then applied to improve the prediction power of the model. The built model can then be used to compute for hazard ratio of two individuals. Likewise, the statistical measure used for the variable selection is the log-likelihood ratio. Table 4.11 presents details of the full model.

Table 4.11: Details of Cox Proportional Hazard with the 23 significant genes

Gene Symbol	coef	se(coef)	z-test	p-value
<i>FIGF</i>	5.71e-03	2.82e-03	2.03	0.0427*
<i>SDPR</i>	1.44e-03	7.37e-04	1.95	0.0507*
<i>CD300LG</i>	-2.43e-03	2.48e-03	-0.98	0.3266
<i>ADAMTS5</i>	-8.18e-04	6.93e-04	-1.18	0.2379
<i>MAMDC2</i>	8.76e-04	7.91e-04	1.11	0.2682
<i>TMEM220</i>	1.20e-03	2.47e-03	0.48	0.6279
<i>SPRY2</i>	-5.04e-04	9.07e-04	-0.56	0.5784
<i>PAMR1</i>	2.38e-03	7.80e-04	3.06	0.0023*
<i>SCARA5</i>	-1.48e-03	1.57e-03	-0.94	0.3474
<i>ARHGAP20</i>	6.56e-04	3.53e-03	0.19	0.8525
<i>TSLP</i>	2.24e-03	9.33e-03	-0.24	0.8101
<i>LMOD1</i>	1.40e-05	4.11e-04	0.03	0.9728
<i>FAM13A</i>	-1.31e-05	4.15e-04	-0.03	0.9749
<i>CA4</i>	2.02e-03	5.13e-03	0.39	0.6934
<i>PPP1R12B</i>	-8.47e-05	2.66e-04	-0.32	0.7499
<i>ABCA10</i>	3.11e-03	4.39e-03	0.71	0.4796
<i>MME</i>	5.51e-05	1.25e-04	0.44	0.6591
<i>SCN4B</i>	-1.83e-03	1.10e-03	-1.67	0.0942*
<i>DMD</i>	-8.32e-04	5.67e-04	-1.47	0.1423
<i>FXYP1</i>	-3.93e-03	3.29e-03	-1.19	0.2326
<i>CAV1</i>	1.57e-04	1.79e-04	0.88	0.3786
<i>ITIH5</i>	-3.86e-04	2.65e-04	-1.45	0.1461
<i>BTNL9</i>	5.35e-04	1.30e-03	0.41	0.6806

“*” means statistically significant

Table 4.11 presents details of a Cox proportion hazard model with 23 genes. The table shows genes with their corresponding coefficients (*coef*), standard error (*se(coef)*), z-test and *p*-value. Again, using a 10% significance level, the model resulted in a global *p*-value equal to 0.0027 with a log-likelihood ratio of 37.8. The results from Table 4.11 shows that *FIGF*, *SDPR*, *PAMR1* and *SCN4B* contribute significantly to the model. To improve on this, the backward elimination

procedure is used to build a new model. Table 4.12 presents the details of the variable selection process.

Table 4.12: Variable Selection Process for Cox Proportional Hazard

Model	Log-likelihood ratio	<i>p</i> -value
model ₁ = Full model	37.8	2.70e-02
model ₂ = model ₁ - FAM13A	37.8	1.95e-02
model ₃ = model ₂ - LMOD1	37.8	1.38e-02
model ₄ = model ₃ - ARHGAP20	37.7	9.58e-03
model ₅ = model ₄ - TSLP	37.7	6.56e-03
model ₆ = model ₅ - PPP1R12B	37.6	4.41e-03
model ₇ = model ₆ - CA4	37.5	2.92e-03
model ₈ = model ₇ - TMEM220	37.3	1.89e-03
model ₉ = model ₈ - MME	37.2	1.20e-03
model ₁₀ = model ₉ - BTNL9	36.9	7.66e-04
model ₁₁ = model ₁₀ - SPRY2	36.5	5.05e-04
model ₁₂ = model ₁₁ - CD300LG	35.8	3.54e-04
model ₁₃ = model ₁₂ - SCARA5	34.9	2.54e-04
model ₁₄ = model ₁₃ - ABCA10	34.5	1.51e-04
model ₁₅ = model ₁₄ - CAV1	33.9	9.32e-05
model ₁₆ = model ₁₅ - FXYD1	32.7	6.99e-05
model ₁₇ = model ₁₆ - ADAMTS5	31.9	4.31e-05

Table 4.12 presents different models with their corresponding log-likelihood ratio and *p*-values. Using Table 4.11, the gene with the largest *p*-value is *FAM13A* (*p*-value for *FAM13A* = 0.9749), hence it is the first to be removed from the model. *ADAMTS5* was the last gene to be removed from the model. Table 4.13, therefore, presents the details of the reduced model.

Table 4.13: Details of the reduced Cox Proportional Hazard

Gene Symbol	coef	se(coef)	z-test	<i>p</i> -value
<i>FIGF</i>	0.003925	0.002066	1.90	5.75e-02*
<i>SDPR</i>	0.001142	0.000244	4.68	2.80e-06*
<i>PAMR1</i>	0.001863	0.000631	2.95	3.10e-03*
<i>SCN4B</i>	-0.002129	0.000952	-2.24	2.54e-02*
<i>DMD</i>	-0.001093	0.000467	-2.34	1.93e-02*
<i>ITIH5</i>	-0.000349	0.000205	-1.70	8.92e-02*
<i>MAMDC2</i>	0.000816	0.000495	1.65	9.88e-02*

* means statistically significant

Table 4.13 shows details of the reduced Cox proportional hazard model. The reduced model contains six genes that contribute significantly to identifying the hazard ratio of patients. Each gene also has its corresponding coefficient (*coef*), standard error (*se(coef)*), z-test and *p*-value. From (3.6.33) the hazard for a breast cancer patients using the Cox proportional hazard is therefore given by

$$h(t) = \exp(0.003925 \text{ FIGF} + 0.001142 \text{ SDPR} + 0.001863 \text{ PAMR1} - 0.002129 \text{ SCN4B} - 0.001093 \text{ DMD} - 0.000349 \text{ ITIH5} + 0.000816 \text{ MAMDC2}). \quad (4.3.5)$$

Equation (4.3.5) illustrates how each gene contributes to the hazard of a patient. For example, if other genes are held constant, an increase of expression of *FIGF* will increase a patient's hazard by a factor of $\exp(0.003925)$; that is by 0.39%. Similarly, a decrease of expression of *SCN4B* will decrease a patient's hazard by a factor of $\exp(-0.002129)$; that is by 0.21%. Again, this is used to evaluate the relative risk of patients by comparing their hazard with a baseline risk. In this case, patients with hazard less than 1 are considered to have a relatively lower risk and patients with hazard more than 1 have a relatively higher risk.

4.4 Discussion of results

Breast cancer is a malignant tumour caused by uncontrolled growth of abnormal cells in the breast. To develop an effective diagnostic and therapeutic tool, the molecular mechanism inherent in breast cancer have to be understood (Wang *et al.*, 2018). RNA-seq is frequently used to profile transcription level within cells which helps to identify genes that are differentially expressed for breast cancer treatment.

In this study, differentially expressed genes for breast cancer patients were identified using SAM. Out of 20532 genes, 23 were identified as significantly different. Specifically, all 23 genes reported by SAM were significantly positive. The study expected genes that were identified as differentially expressed to make biological sense, hence the functions of the genes are investigated.

FIGF which was ranked by SAM as the most significantly positive gene functions as a growth factor in a cell. It is mainly involved in angiogenesis, which is the process of forming new blood vessels (Hayes, 1994; Schneider and Miller,

2005). Tian *et al.* (2016) discovered that *SDPR* inhibits the progression of breast cancer and this may be due to the fact that *SDPR* starves the breast cancer cells from serum. This leads to it suppressing breast cancer cell proliferation and invasion. Their further investigation revealed *SDPR* to be down-regulated in human breast cancer. Shen *et al.* (2015) also revealed *CD300LG* to be down-regulated in breast cancer tumours.

Furthermore, Nissinen and Khri (2012) reported that *ADAMTS5* may function as a tumour suppressor because they oppose the growth of a tumour and demonstrate antagonistic behaviour to angiogenesis. They further found evidence that *ADAMTS5* was down-regulated in malignant tumour progression. In addition, a study by Porter *et al.* (2004) revealed that *ADAMTS5* was down-regulated in human breast cancer. Tishchenko *et al.* (2016) discovered *MAMDC2* to be associated with tumour necrosis and were down-regulated in breast cancer cells. Also, Choi *et al.* (2017) identified *TMEM220* as a down-regulated gene in gastric cancer. Feng *et al.* (2011) found *SPRY2* to contribute to tumorigenesis when they are deregulated. Faratian *et al.* (2011) also describe *SPRY2* to function as a tumour suppressor.

Lo *et al.* (2015) stated *PAMR1* as a putative tumour suppressor as they suppress the growth of cancer cells. They then noted *PAMR1* to be suppressed in breast cancer cells. Ulker *et al.* (2018) found that *SCARA5* had significantly decreased in cancerous tissues compared to that of non-cancerous samples. They discovered that the down-regulation was associated with hypermethylation of the promoter and thus plays an important role in tumorigenesis. Yamada *et al.* (2005) predicted *ARHGAP20* to be a tumour suppressor gene activated by deletion in breast cancer.

TSLP blockade could be an important therapy for cancer Lokuan and Ziegler (2014) because they promote the survival of tumour cells through induction of the expression of an anti-apoptotic molecule (Kuan and Ziegler, 2018). According to Guo *et al.* (2015), *LMOD1* are mostly involved in smooth muscle functions. They also used SAMseq to identify *LMOD1* as an upregulated gene. *FAM13A* are most highly expressed in lung cancer. They are mainly induced by hypoxia effect to reduce the amount of oxygen distributed to tumour cells. Their functions are therefore involved in cell proliferation. Their expression in breast cancer tissues may be attributed to the fact that, the breast lays close to the lung (Eisenhut

et al., 2017; Ziłkowska-Suchanek *et al.*, 2017).

Similarly, *CA4* mostly expressed in lung cancer cells (Carter *et al.*, 1990). Again, Lo *et al.* (2015) identified *PPP1R12B* to be frequently down-regulated in breast cancer tissues. *ABCA10* is a lipid transporter but proved to correlate with breast cancer (Wang *et al.*, 2015). *MME* was identified to be highly expressed in breast cancer progression and dissemination (Louhichi *et al.*, 2018).

Bon *et al.* (2016) reported that *SCN4B* is a tumour suppressor gene and reduces cancer cell invasiveness and tumour progression. *DMD* has been validated as a new agent in tumour development for tumour progression (Luce *et al.*, 2017). Zhang *et al.* (2013) stated that supplementing oestrogen deficient can result in complications such as breast cancer. However, *FXYD1* can be used to inhibit the expression of miR-151-5p which is associated with oestrogen deficiency. *CAV1* has been discovered to function as a primary tumour growth regulator (Sloan *et al.*, 2004). It, therefore, plays an important role as a tumour suppressor in breast cancer cells and is a therapeutic target for the treatment of breast cancer (Mercier and Lisanti, 2012). *ITIH5* was recently identified to impair breast cancer progression but its underlying functions are still unclear (Rose *et al.*, 2017). It is however associated to be a tumour suppressor as its absence increases the rate of proliferation (Veeck *et al.*, 2008). Hsu *et al.* (2017) identified *BTNL9* to function as a tumour suppressor in lung cancer.

Summarising the literature discussed above, 8 genes were found to be associated with tumour suppression and they are *SDPR*, *ADAMTS5*, *PAMR1*, *ARHGAP20*, *SCN4B*, *CAV1*, *ITIH5*, and *BTNL9*. The study noted that when they are down-regulated, breast cancer is promoted. *FIGF* is both associated with angiogenesis and cell growth factor. Both *SPRY2* and *SCARA5* contributed to tumorigenesis when they are down-regulated. There were 4 antagonist genes that rather promotes breast cancer when up-regulated. They are *FAM13A*, *TSLP*, *MME* and *DMD*. When up-regulated, *FAM13A* contributes to cell proliferation, *TSLP* also contributes to tumour survival, and both *MME* and *DMD* contribute to tumour progression. Only *MAMDC2* gene was associated with tumour necrosis when it was down-regulated. Also, *CD300LG* and *PPP1R12B* are only down-regulated in breast cancer. *TMEM220* and *CA4* are related to gastric cancer and lung cancer respectively.

Next, the study compared three different classifiers namely; random forest, artificial neural network and support vector machine. The comparison was performed on three different datasets. In the first data, the two classes –which is tumour class and tumour-free class– were not balanced. The second and third data were then balanced using upsampling techniques called median supplement and mean supplement respectively. After using a confusion matrix as a measure to compare the performance of the three classifiers, random forest performed better than the other classifiers in all three datasets. The performance of random forest may be attributed to its bagging property which helps to decorrelate variables to reduce variance and to also overcome overfitting (Friedman *et al.*, 2001). Hence, the results from random forests may be likened to a combined decision obtained from a group of doctors based on a patient's gene expression data. Moreover, when the random forest was investigated further, it suggested *FIGF*, *ADAMTS5*, *SPRY2*, *ARHGAP20*, *CD300LG*, *SCN4B*, *TMEM220*, *MME*, *SDPR*, and *MAMDC2* as prognostic factors that increase the prediction of breast cancer. In Figure 4.3, the model was used to grow a tree. The tree grown is relatively small, therefore, clinicians can quickly make diagnoses for further treatment of the disease.

Finally, the study investigated how genes influence the survival of patients. It started by using Kaplan–Meier survival curves to estimate the median survival time for breast cancer patients. The results indicated that averagely, a breast cancer patient will survive for 3472 days. In addition, the study used the log–rank test statistic to identify *SDPR*, *ADAMTS5*, *TMEM220*, *SPRY2*, *PAMR1*, *TSLP*, *CAV1* and *BTNL9* as independent prognostic factors to the survival of breast cancer patients. Prognostic factors identified may be attributed to the fact that the survival time of patients expressing more of the prognostic factors was significantly different from the survival time of patients expression less of these prognostic factors.

Further investigations revealed that log–logistic distribution best describes the survival time for breast cancer patients. This was because log–logistic distribution had the smallest AIC value of 2050.11. This means that among the probability distributions log–logistic relatively provides the least information lost from the patient's survival time. From Figure 4.7, it is seen that log–logistic distribution was very close to the curve estimated non–parametrically by Kaplan–Meier.

The study does not use the Kaplan–Meier curve because the study is interested in building a survival curve which incorporates significantly identified genes as prognostic factors. However, the Kaplan–Meier provides a picture of the expected survival curve.

Moreover, the study built a survival model using the log-logistic distribution to predict the hazard and survival probability of breast cancer patients. A significant model containing all 23 genes identified by SAM was first built with a p -value of 0.019. The model was then improved using a back elimination variable selection. This yielded a new p -value of 0.000025. Genes included in the improved model were *FIGF*, *SDPR*, *PAMR1*, *SCN4B*, *DMD*, and *ITIH5*. Specifically, when *FIGF*, *SDPR* and *PAMR1* are increased, the hazard of breast cancer patients decreases by a factor of 0.23%, 0.07% and 0.12% respectively. Whereas when *SCN4B*, *DMD* and *ITIH5* are increased, the hazard of breast cancer patients increases by a factor of 0.12%, 0.05% and 0.02% respectively. Hence, these genes may serve as a potential prognostic factor to determine breast cancer patients survival details. In addition, Figure 4.8 suggests that breast cancer patients at the beginning of the disease will have an increased risk to the event but after 3243.38 days, their risk to the event may gradually decrease.

Again, the study employed Cox proportional hazard to investigate the relative risk of breast cancer patients. Similarly, backward elimination was used to improve a significant model which contained all 23 genes identified by SAM. The improved model contained 7 genes namely; *FIGF*, *SDPR*, *PAMR1*, *SCN4B*, *DMD*, *MAMDC2* and *ITIH5*. It can be seen in Equation (4.3.2) that Cox proportional hazard contained all 6 genes in the log–logistic survival model and *MAMDC2*. Similarly, the Cox proportional hazard model suggests that, when *FIGF*, *SDPR*, *PAMR1* and *MAMDC2* are increased, the hazard of breast cancer patients decreases by a factor of 0.39%, 0.11%, 0.19% and 0.08% respectively. While, *SCN4B*, *DMD* and *ITIH5* increase the hazard of breast cancer patients by a factor of 0.21%, 0.10% and 0.03% when they are increased.

These results suggest that there exists a correlation between the regulation of genes and risk to the event. In particular, it is identified that all genes which promote breast cancer when down–regulated had a negative coefficient in the survival model. Hence, genes that promote breast cancer when down–regulated reduce patient’s risk to the event when they are increased. Similarly, the study

identified that genes which promote breast cancer when upregulated had a positive coefficient in the survival model. Thus, genes that promote breast cancer when upregulated increases patient's risk to the event when they are increased.

In summary, this chapter used SAM to identify genes that are differentially expressed in breast cancer patients. The identified genes were then used to predict breast cancer using three machine learning algorithms. The study then used a confusion matrix to compare the performance of three machine learning algorithms. Finally, the significant genes were used to build a model to predict the survival time of breast cancer patients.

Chapter 5

Conclusion

In conclusion, the study identified potential genes that may contribute to the formation and development of the breast cancer disease. This was achieved by using a statistical technique called significance analysis of microarray (SAM) to identify significant genes. Identified significant genes were used as features of supervised learning techniques to predict breast cancer and a survival model to predict the survival probability of breast cancer in patients.

In the experimental investigation, we identified 23 significantly positive genes to be differentially expressed in breast cancer patients (see Table 4.1). Further investigation revealed that most positive genes when down-regulated promote the formation of breast cancer. This may lead to discovering novel genes that contribute to breast cancer and enable clinicians to target specific genes for treating breast cancer.

Moreover, the study also concludes that random forest is well suited for predicting breast cancer with *FIGF*, *SDPR*, *CD300LG*, *ADAMTS5*, *MAMDC2*, *TMEM220*, *SPRY2*, *ARHGAP20*, *MME* and *SCN4B* as predictive variables. The model was used to construct a simple tree, which is easily interpretable and may be used by clinicians to diagnose breast cancer (See Figure 4.3).

Finally, the study identified the median survival time for breast cancer patients to be 3472 days. Further analysis using probability distributions commend using log-logistic distribution to build a model to predict survival probability for breast cancer patients with *FIGF*, *SDPR*, *PAMR1*, *SCN4B*, *DMD* and *ITIH5* as prognostic factors. The survival model further revealed that breast cancer pa-

tients will have an increased risk to the event but after 3243.38 days, their risk to the event may gradually reduce.

In the future, we intend to create a gene–gene interaction network to discover functional relationships that exist between identified genes at the systems level. Lastly, we intend to use random survival forest to build a model to uncover interrelationship which exists between predictive variables and prognostic factors of breast cancer ([Ishwaran *et al.*, 2008](#)).

List of references

- Abadi, A., Yavari, P., Dehghani-Arani, M., Alavi-Majd, H., Ghasemi, E., Amanpour, F. and Bajdik, C. (2014). Cox models survival analysis based on breast cancer treatments. *Iranian journal of cancer prevention*, vol. 7, no. 3, p. 124.
- Adabor, E.S. and Acquaaah-Mensah, G.K. (2017). Machine learning approaches to decipher hormone and HER2 receptor status phenotypes in breast cancer. *Briefings in bioinformatics*, , no. June 2017, pp. 1–11. ISSN 1477-4054 (Electronic).
- Anders, S. and Huber, W. (2010). Differential expression analysis for sequence count data. *Genome biology*, vol. 11, no. 10, p. R106.
- Bishop, C.M. (2016). *Pattern Recognition and Machine Learning*. Information Science and Statistics. Springer New York. ISBN 9781493938438.
- Bon, E., Driffort, V., Gradek, F., Martinez-Caceres, C., Anchelin, M., Pelegrin, P., Cayuela, M.-L., Marionneau-Lambot, S., Oullier, T., Guibon, R., Fromont, G., Gutierrez-Pajares, J.L., Domingo, I., Piver, E., Moreau, A., Burlaud-Gaillard, J., Frank, P.G., Chevalier, S., Besson, P. and Roger, S. (2016). SCN4B acts as a metastasis-suppressor gene preventing hyperactivation of cell migration in breast cancer. *Nature communications*, vol. 7, p. 13648.
- Bottou, L. and Lin, C. (2007). Support vector machine solvers. *Large-scale kernel machines*. Available at: [http://140.112.30.28/\\$\sim\\$cjlin/papers/bottou_lin.pdf](http://140.112.30.28/\simcjlin/papers/bottou_lin.pdf)
- Breiman, L. (2001). Random forests. *Machine learning*, vol. 45, no. 1, pp. 5–32.
- Brown, M.P.S., Grundy, W.N., Lin, D., Cristianini, N., Sugnet, C.W., Furey, T.S., Jr and Haussler, D. (2000). Knowledge-based analysis of microarray gene expression data by using support vector machines. *Pnas*, vol. 97, no. 1, pp. 262–267.

- Carey, L.A., Perou, C.M., Livasy, C.A., Dressler, L.G., Cowan, D., Conway, K., Karaca, G., Troester, M.A., Tse, C.K., Edmiston, S. *et al.* (2006). Race, breast cancer subtypes, and survival in the Carolina Breast Cancer Study. *Jama*, vol. 295, no. 21, pp. 2492–2502.
- Carter, N., Fryer, A., Grant, A., Hume, R., Strange, R. and Wistrand, P. (1990). Membrane specific carbonic anhydrase (CAIV) expression in human tissues. *Biochimica et Biophysica Acta (BBA)-Biomembranes*, vol. 1026, no. 1, pp. 113–116.
- Chai, H., Li, Z.-N., Meng, D.-Y., Xia, L.-Y. and Liang, Y. (2017). A new semi-supervised learning model combined with Cox and SP-AFT models in cancer survival analysis. *Scientific reports*, vol. 7, no. 1, p. 13053.
- Chen, X., Sun, X. and Hoshida, Y. (2014). Survival analysis tools in genomics research. *Human genomics*, vol. 8, no. 1, p. 21.
- Choi, B., Han, T.-S., Min, J., Hur, K., Lee, S.-M., Lee, H.-J., Kim, Y.-J. and Yang, H.-K. (2017). MAL and TMEM220 are novel DNA methylation markers in human gastric cancer. *Biomarkers*, vol. 22, no. 1, pp. 35–44.
- Chu, G., Li, J., Narasimhan, B., Tibshirani, R. and Tusher, V. (2001). Significance analysis of microarrays users guide and technical document.
- Clamp, M., Fry, B., Kamal, M., Xie, X., Cuff, J., Lin, M.F., Kellis, M., Lindblad-Toh, K. and Lander, E.S. (2007). Distinguishing protein-coding and noncoding genes in the human genome. *Proceedings of the National Academy of Sciences*, vol. 104, no. 49, pp. 19428–19433.
- Collett, D. (1993). *Modelling survival data in medical research*. Chapman and Hall/CRC.
- Corney, D.C. and Basturea, N.G. (2013). RNA-seq using next generation sequencing. *Materials and Methods*, vol. 3, p. 203.
- Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, vol. 20, no. 3, pp. 273 – 297.
- Costa-Silva, J., Domingues, D. and Lopes, F.M. (2017). RNA-Seq differential expression analysis: An extended review and a software tool. *PloS one*, vol. 12, no. 12, p. e0190152.
- Danaee, P., Ghaeini, R. and Hendrix, D.A. (2017). A Deep Learning Approach for Cancer Detection and Relevant Gene Identification. *Biocomputing 2017*, pp. 219–229.

- Douglas, P.K., Harris, S., Yuille, A. and Cohen, M.S. (2011). Performance comparison of machine learning algorithms and number of independent components used in fMRI decoding of belief vs. disbelief. *Neuroimage*, vol. 56, no. 2, pp. 544–553.
- Draghici, S., Khatri, P., Eklund, A.C. and Szallasi, Z. (2006). Reliability and reproducibility issues in DNA microarray measurements. *TRENDS in Genetics*, vol. 22, no. 2, pp. 101–109.
- Efron, B. and Hastie, T. (2016). *Computer age statistical inference*, vol. 5. Cambridge University Press.
- Eisenhut, F., Heim, L., Trump, S., Mittler, S., Sopel, N., Andreev, K., Ferrazzi, F., Ekici, A.B., Rieker, R., Springel, R., Assmann, V.L., Lechmann, M., Koch, S., Engelhardt, M., Warnecke, C., Trufa, D.I., Sirbu, H., Hartmann, A. and Finotto, S. (2017). FAM13A is associated with non-small cell lung cancer (NSCLC) progression and controls tumor cell proliferation and survival. *OncImmunity*, vol. 6, no. 1, pp. 1–15.
- Ezkurdia, I., Juan, D., Rodriguez, J.M., Frankish, A., Diekhans, M., Harrow, J., Vazquez, J., Valencia, A. and Tress, M.L. (2014). Multiple evidence strands suggest that there may be as few as 19000 human protein-coding genes. *Human molecular genetics*, vol. 23, no. 22, pp. 5866–5878.
- Faratian, D., Sims, A.H., Mullen, P., Kay, C., Um, I.H., Langdon, S.P. and Harrison, D.J. (2011). Sprouty 2 is an independent prognostic factor in breast cancer and may be useful in stratifying patients for trastuzumab therapy. *PLoS ONE*, vol. 6, no. 8.
- Faraway, J.J. (2002). *Practical Regression and ANOVA using R*.
- Feng, Y.-H., Wu, C.-L., Tsao, C.-J., Chang, J.-G., Lu, P.-J., Yeh, K.-T., Uen, Y.-H., Lee, J.-C. and Shiau, A.-L. (2011). Deregulated expression of sprouty2 and microRNA-21 in human colon cancer: Correlation with the clinical stage of the disease. *Cancer Biology & Therapy*, vol. 11, no. 1, pp. 111–121. ISSN 1538-4047.
- Friedman, J., Hastie, T. and Tibshirani, R. (2001). *The elements of statistical learning*, vol. 1. Springer series in statistics New York, NY, USA:.
- Glander, S. (2018 August). Dealing with unbalanced data in machine learning. shiring.github.io.
Available at: https://shiring.github.io/machine_learning/2017/04/02/unbalanced

- Goodfellow, I., Bengio, Y. and Courville, A. (2016). *Deep learning*, vol. 1. MIT press Cambridge.
- Guo, X., Jo, V.Y., Mills, A.M., Zhu, S.X., Lee, C.-H., Espinosa, I., Nucci, M.R., Varma, S., Forgó, E., Hastie, T. *et al.* (2015). Clinically relevant molecular subtypes in leiomyosarcoma. *Clinical cancer research*.
- Hardcastle, T.J. and Kelly, K.A. (2010). baySeq: empirical Bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, vol. 11, no. 1, p. 422.
- Hayes, D.F. (1994). Angiogenesis and breast cancer. *Hematology/Oncology Clinics*, vol. 8, no. 1, pp. 51–71.
- Hsu, Y.-L., Hung, J.-Y., Lee, Y.-L., Chen, F.-W., Chang, K.-F., Chang, W.-A., Tsai, Y.-M., Chong, I.-W. and Kuo, P.-L. (2017). Identification of novel gene expression signature in lung adenocarcinoma by using next-generation sequencing data and bioinformatics analysis. *Oncotarget*, vol. 8, no. 62, p. 104831.
- Ishwaran, H., Kogalur, U.B., Blackstone, E.H., Lauer, M.S. *et al.* (2008). Random survival forests. *The annals of applied statistics*, vol. 2, no. 3, pp. 841–860.
- Jean-Philippe Vert (2001). Introduction to Support Vector Machines and Applications to Computational Biology.
- Jiangeng, L., Yanhua, D. and Xiaogang, R. (2007). A novel hybrid approach to selecting marker genes for cancer classification using gene expression data. In: *Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on*, pp. 264–267. IEEE.
- Kaplan, E.L. and Meier, P. (1958). Nonparametric Estimation from Incomplete Observations. *Journal of the American Statistical Association*, vol. 53, no. 282, pp. 457–481.
- Karpathy, A. (2017 June). CS231n Convolutional Neural Networks for Visual Recognition.
Available at: <http://cs231n.github.io/neural-networks-1/#nn>
- Khondoker, M., Dobson, R., Skirrow, C., Simmons, A. and Stahl, D. (2016). A comparison of machine learning methods for classification using simulation with multiple real data examples from mental health studies. *Statistical methods in medical research*, vol. 25, no. 5, pp. 1804–1823.

- Kleinbaum, D.G. (1998). Survival Analysis, a Self-Learning Text. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, vol. 40, no. 1, pp. 107–108.
- Kourou, K., Exarchos, T.P., Exarchos, K.P., Karamouzis, M.V. and Fotiadis, D.I. (2015). Machine learning applications in cancer prognosis and prediction. *Computational and Structural Biotechnology Journal*, vol. 13, pp. 8–17. ISSN 20010370.
- Kuan, E.L. and Ziegler, S.F. (2018). Publisher Correction: A tumor–myeloid cell axis, mediated via the cytokines IL-1 α and TSLP, promotes the progression of breast cancer. *Nature Immunology*, p. 1. ISSN 15292916.
- LeCun, Y.A., Bottou, L., Orr, G.B. and Müller, K.-R. (2012). Efficient backprop. In: *Neural networks: Tricks of the trade*, pp. 9–48. Springer.
- Lee, E. and Wang, J.W. (2003). *Statistical methods for survival data analysis*, vol. 36.
- Li, B. and Dewey, C.N. (2011). RSEM: Accurate transcript quantification from RNA-Seq data with or without a reference genome. *BMC Bioinformatics*, vol. 12. ISSN 14712105.
- Li, B., Ruotti, V., Stewart, R.M., Thomson, J.A. and Dewey, C.N. (2009). RNA-Seq gene expression estimation with read mapping uncertainty. *Bioinformatics*, vol. 26, no. 4, pp. 493–500. ISSN 14602059.
- Li, J. and Tibshirani, R. (2013). Finding consistent patterns: a nonparametric approach for identifying differential expression in RNA-Seq data. *Statistical methods in medical research*, vol. 22, no. 5, pp. 519–536.
- Li, J., Witten, D.M., Johnstone, I.M. and Tibshirani, R. (2012). Normalization, Testing, and False Discovery Rate Estimation for RNA-Sequencing Data. *Biostatistics*, vol. 13, no. 3, pp. 523–538.
- Li, S.Q., Pan, X.F., Kashaf, M.S., Xue, Q.P., Luo, H.J., Wang, Y.Y., Wen, Y. and Yang, C.X. (2017). Five-Year Survival is Not a Useful Measure for Cancer Control in the Population: an Analysis Based on UK Data. *Asian Pacific journal of cancer prevention: APJCP*, vol. 18, no. 2, p. 571.
- Li, W.V. and Li, J.J. (2018). Modeling and analysis of RNA-seq data: a review from a statistical perspective. *arXiv preprint arXiv:1804.06050*.
- Liang, Y., Chai, H., Liu, X.-Y., Xu, Z.-B., Zhang, H. and Leung, K.-S. (2016). Cancer survival analysis using semi-supervised learning method based on Cox and AFT models with $L_{1/2}$ regularization. *BMC medical genomics*, vol. 9, no. 1, p. 11.

- Liaw, A. and Wiener, M. (2002). Classification and Regression by randomForest. *R news*, vol. 2, no. December, pp. 18–22. ISSN 16093631.
- Lo, P.H.Y., Tanikawa, C., Katagiri, T., Nakamura, Y. and Matsuda, K. (2015). Identification of novel epigenetically inactivated gene PAMR1 in breast carcinoma. *Oncology Reports*, vol. 33, no. 1, pp. 267–273. ISSN 17912431.
- Lokuan, E. and Ziegler, S.F. (2014). Thymic stromal lymphopoietin (TSLP) and cancer. *Journal of Immunology*, vol. 193, no. 9, pp. 4283–4288. ISSN 1946-6242.
- Louhichi, T., Saad, H., Dhiab, M.B., Ziadi, S. and Trimeche, M. (2018). Stromal CD10 expression in breast cancer correlates with tumor invasion and cancer stem cell phenotype. *BMC cancer*, vol. 18, no. 1, p. 49.
- Love, M.I., Huber, W. and Anders, S. (2014). Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology*, vol. 15, no. 12, pp. 1–21.
- Lowe, R., Shirley, N., Bleackley, M., Dolan, S. and Shafee, T. (2017). Transcriptomics technologies. *PLoS Computational Biology*, vol. 13, no. 5, pp. 1–23. ISSN 15537358.
- Luce, L.N., Abbate, M., Cotignola, J. and Giliberto, F. (2017). Non-myogenic tumors display altered expression of dystrophin (DMD) and a high frequency of genetic alterations. *Oncotarget*, vol. 8, no. 1, p. 145.
- McGready, J. (2009). Regression for Survival Analysis.
Available at: http://ocw.jhsph.edu/courses/StatisticalReasoning2/PDFs/2009/StatR2_lec10a_mcgready.pdf
- McKay, M.D., Beckman, R.J. and Conover, W.J. (1979). Comparison of three methods for selecting values of input variables in the analysis of output from a computer code. *Technometrics*, vol. 21, no. 2, pp. 239–245.
- Mensah, S.O., Mazandu, G.K. and Utete, S.W. (2017). *Investigating Relationship Between Expressed Cancer Related Genes and Patient Survival*. Master's thesis, African Institute for Mathematical Sciences (AIMS).
- Mercier, I. and Lisanti, M.P. (2012). Caveolin-1 and breast cancer: a new clinical perspective. In: *Caveolins and Caveolae*, pp. 83–94. Springer.
- Miecznikowski, J.C., Wang, D., Liu, S., Sucheston, L. and Gold, D. (2010). Comparative survival analysis of breast cancer microarray studies identifies important prognostic genetic pathways. *BMC cancer*, vol. 10, no. 1, p. 573.

- Nasejje, J. (2012). *Statistical Methods Used in Survival Analysis*. Master's thesis, African Institute for Mathematical Sciences (AIMS).
- National Cancer Institute (2017 April). Genetics.
Available at: <https://cancer.gov/about-cancer/causes-prevention/genetics>
- Nissinen, L. and Khri, V.M. (2012). ADAMTS5: A new player in the vascular field. *American Journal of Pathology*, vol. 181, no. 3, pp. 743–745. ISSN 00029440.
- Polat, K. and Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital signal processing*, vol. 17, no. 4, pp. 694–701.
- Porter, S., Scott, S.D., Sassoon, E.M., Williams, M.R., Jones, J.L., Girling, A.C., Ball, R.Y. and Edwards, D.R. (2004). Dysregulated Expression of Adamalysin-Thrombospondin Genes in Human Breast Carcinoma Dysregulated Expression of Adamalysin-Thrombospondin Genes in Human Breast Carcinoma. pp. 2429–2440.
- Poulin, N.M. and Nielsen, T.O. (2009). Expression arrays: Discovery and Validation. In: *Cell and Tissue based Molecular Pathology*, pp. 70–83. Elsevier.
- Raschka, S. (2015). *Python machine learning*. Packt Publishing Ltd.
- Robinson, M.D., McCarthy, D.J. and Smyth, G.K. (2009). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, vol. 26, no. 1, pp. 139–140. ISSN 14602059.
- Rodriguez, G. (2007). Survival Models. *Lecture Notes on Generalized Linear Models*, , no. 1.
Available at: <http://data.princeton.edu/wws509/notes/c7.pdf>
- Rose, M., Kloten, V., Noetzel, E., Gola, L., Ehling, J., Heide, T., Meurer, S.K., Gaikoshcherbak, A., Sechi, A.S., Huth, S. *et al.* (2017). ITIH5 mediates epigenetic reprogramming of breast cancer cells. *Molecular cancer*, vol. 16, no. 1, p. 44.
- Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986). Learning representations by back-propagating errors. *nature*, vol. 323, no. 6088, p. 533.
- Schneider, B.P. and Miller, K.D. (2005). Angiogenesis of breast cancer. *Journal of Clinical Oncology*, vol. 23, no. 8, pp. 1782–1790.
- Seyednasrollah, F., Laiho, A. and Elo, L.L. (2013). Comparison of software packages for detecting differential expression in RNA-seq studies. *Briefings in Bioinformatics*, vol. 16, no. 1, pp. 59–70. ISSN 14774054.

- Shen, X., Xie, B., Ma, Z., Yu, W., Wang, W., Xu, D., Yan, X., Chen, B., Yu, L., Li, J., Chen, X., Ding, K. and Cao, F. (2015). Identification of novel long non-coding RNAs in triple-negative breast cancer. *Oncotarget*, vol. 6, no. 25, pp. 21730–21739.
- Siegel, R.L., Miller, K.D. and Jemal, A. (2018). Cancer statistics, 2018. *CA: A Cancer Journal for Clinicians*, vol. 68, no. 1, pp. 7–30.
- Sloan, E.K., Stanley, K.L. and Anderson, R.L. (2004). Caveolin-1 inhibits breast cancer growth and metastasis. *Oncogene*, vol. 23, no. 47, p. 7893.
- Soneson, C. and Delorenzi, M. (2013). A comparison of methods for differential expression analysis of RNA-seq data. *BMC Bioinformatics*, vol. 14. ISSN 14712105.
- Stein, M. (1987). Large sample properties of simulations using Latin hypercube sampling. *Technometrics*, vol. 29, no. 2, pp. 143–151.
- Strobl, C., Malley, J. and Gerhard Tutz (2009). Characteristics of Classification and Regression Trees, Bagging and Random Forests. *Psychol Methods*, vol. 14, no. 4, pp. 323–348.
- Su, Z., Łabaj, P.P., Li, S., Thierry-Mieg, J., Thierry-Mieg, D., Shi, W., Wang, C., Schroth, G.P., Setterquist, R.A., Thompson, J.F. *et al.* (2014). A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nature biotechnology*, vol. 32, no. 9, p. 903.
- Teknomo, K. (2009). Tutorial on Decision Tree.
Available at: <http://people.revoledu.com/kardi/tutorial/DecisionTree/>
- Tian, Y., Yu, Y., Hou, L.K., Chi, J.R., Mao, J.F., Xia, L., Wang, X., Wang, P. and Cao, X.C. (2016). Serum deprivation response inhibits breast cancer progression by blocking transforming growth factor- β signaling. *Cancer Science*, vol. 107, no. 3, pp. 274–280.
- Tishchenko, I., Milioli, H.H., Riveros, C. and Moscato, P. (2016). Extensive transcriptomic and genomic analysis provides new insights about luminal breast cancers. *PLoS ONE*, vol. 11, no. 6, pp. 1–36. ISSN 19326203.
- Tomczak, K., Czerwińska, P. and Wiznerowicz, M. (2015). The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemporary oncology*, vol. 19, no. 1A, p. A68.
- Tong, S. and Koller, D. (2000). Restricted Bayes optimal classifiers. *Proceedings of the National Conference on Artificial Intelligence*, pp. 658–664.

- Tusher, V.G., Tibshirani, R. and Chu, G. (2001). Significance Analysis of Microarrays Applied to the Ionizing Radiation Response. *Proceedings of the National Academy of Sciences*, vol. 98, no. 9, pp. 5116–5121.
- Ulker, D., Ersoy, Y.E., Gucin, Z., Muslumanoglu, M. and Buyru, N. (2018). Downregulation of SCARA5 may contribute to breast cancer via promoter hypermethylation. *Gene*, vol. 673, no. 2017, pp. 102–106. ISSN 18790038.
- van 't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A.M., Mao, M., Peterse, H.L., van der Kooy, K., Marton, M.J., Witteveen, A.T., Schreiber, G.J., Kerkhoven, R.M., Roberts, C., Linsley, P.S., Bernardis, R. and Friend, S.H. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, vol. 415, no. 6871, pp. 530–536. ISSN 00280836. [415530a](https://doi.org/10.1038/415530a).
Available at: <http://www.nature.com/doi/10.1038/415530a>
- Vanneschi, L., Farinaccio, A., Mauri, G., Antoniotti, M., Provero, P. and Giacobini, M. (2011). A comparison of machine learning techniques for survival prediction in breast cancer. *BioData Mining*, vol. 4, no. 1, p. 12. ISSN 1756-0381.
- Veeck, J., Chorovicer, M., Naami, A., Breuer, E., Zafrakas, M., Bektas, N., Dürst, M., Kristiansen, G., Wild, P., Hartmann, A. *et al.* (2008). The extracellular matrix protein ITIH5 is a novel prognostic marker in invasive node-negative breast cancer and its aberrant expression is caused by promoter hypermethylation. *Oncogene*, vol. 27, p. 865.
- Wang, H.-J., Zhou, M., Jia, L., Sun, J., Shi, H.-B., Liu, S.-L. and Wang, Z.-Z. (2015). Identification of aberrant chromosomal regions in human breast cancer using gene expression data and related gene information. *Medical science monitor: international medical journal of experimental and clinical research*, vol. 21, p. 2557.
- Wang, Y., Zhang, Y., Huang, Q. and Li, C. (2018). Integrated bioinformatics analysis reveals key candidate genes and pathways in Breast cancer. *Molecular Medicine Reports*, vol. 17, no. 6, pp. 8091–8100. ISSN 17913004.
- Wang, Z., Gerstein, M. and Snyder, M. (2009). RNA-Seq: a revolutionary tool for transcriptomics. *Nature reviews genetics*, vol. 10, no. 1, p. 57.
- Wolpert, D.H. and Macready, W.G. (1995). No free lunch theorems for search. Tech. Rep., Technical Report SFI-TR-95-02-010, Santa Fe Institute.

- World Health Organisation (2017 February). Cancer. <http://www.who.int/mediacentre/factsheets/fs297/en/>. Accessed: April 23, 2017.
- Yamada, T., Sakisaka, T., Hisata, S., Baba, T. and Takai, Y. (2005). RA-RhoGAP, Rap-activated Rho GTPase-activating protein implicated in neurite outgrowth through Rho. *Journal of Biological Chemistry*, vol. 280, no. 38, pp. 33026–33034. ISSN 00219258.
- Yue, W., Wang, Z., Chen, H., Payne, A. and Liu, X. (2018). Machine Learning with Applications in Breast Cancer Diagnosis and Prognosis. *Designs*, vol. 2, no. 2, p. 13.
- Zhang, Y., Wang, R., Du, W., Wang, S., Yang, L., Pan, Z., Li, X., Xiong, X., He, H., Shi, Y. *et al.* (2013). Downregulation of mir-151-5p contributes to increased susceptibility to arrhythmogenesis during myocardial infarction with estrogen deprivation. *PloS one*, vol. 8, no. 9, p. e72985.
- Zhang, Z. (2016). Semi-parametric regression model for survival data: graphical visualization with R. *Annals of Translational Medicine*, vol. 4, no. 23, pp. 461–461.
- Zhang, Z.H., Jhaveri, D.J., Marshall, V.M., Bauer, D.C., Edson, J., Narayanan, R.K., Robinson, G.J., Lundberg, A.E., Bartlett, P.F., Wray, N.R. and Zhao, Q.Y. (2014). A comparative study of techniques for differential expression analysis on RNA-seq data. *PLoS ONE*, vol. 9, no. 8. ISSN 19326203.
- Zhao, S., Fung-Leung, W.-P., Bittner, A., Ngo, K. and Liu, X. (2014). Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one*, vol. 9, no. 1, p. e78644.
- Ziółkowska-Suchanek, I., Mosor, M., Podralska, M., Izykowska, K., Gabryel, P., Dyszkiewicz, W., Słomski, R. and Nowak, J. (2017). FAM13A as a novel hypoxia-induced gene in non-small cell lung cancer. *Journal of Cancer*, vol. 8, no. 19, pp. 3933–3938. ISSN 18379664.