



Development of Artificial Intelligence systems as a prediction tool in ovarian cancer

Amir Enshaei

Thesis submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy
Newcastle University
Faculty of Medical Sciences
Northern Institute for Cancer Research

June 2012

Declaration

I certify that no part of the material documented in this thesis has previously been submitted for a degree or other qualification in this or any other university. I declare that this thesis represents my own unaided work, carried out by myself, except where it is acknowledged otherwise in the thesis text.

Amir Enshaei

January 2012

Abstract

Ovarian cancer is the 5th most common cancer in females and the UK has one of the highest incident rates in Europe. In the UK only 36% of patients will live for at least 5 years after diagnosis. The number of prognostic markers, treatments and the sequences of treatments in ovarian cancer are rising. Therefore, it is getting more difficult for the human brain to perform clinical decision making. There is a need for an expert computer system (e.g. Artificial Intelligence (AI)), which is capable of investigating the possible outcomes for each marker, treatment and sequence of treatment. Such expert systems may provide a tool which could help clinicians to analyse and predict outcome using different treatment pathways.

Whilst prediction of overall survival of a patient is difficult there may be some benefits, as this not only is useful information for the patient but may also determine treatment modality.

In this project a dataset was constructed of 352 patients who had been treated at a single centre. Clinical data were extracted from the health records. Expert systems were then investigated to determine the optimum model to predict overall survival of a patient. The five year survival period (a standard survival outcome measure in cancer research) was investigated; in addition, the system was developed with the flexibility to predict patient survival rates for many other categories. Comparisons with currently used prognostic models in ovarian cancer demonstrated a significant improvement in performance for the AI model (Area under the Curve (AUC) of 0.72 for AI and AUC of 0.62 for the statistical model). Using various methods, the most important variables in this prediction were identified as: FIGO stage, outcome of the surgery and CA125. This research investigated the effects of increasing the number of cases in prediction models. Results indicated that by increasing the number of cases, the prediction performance improved. Categorization of continuous data did not improve the prediction performance.

The project next investigated the possibility of predicting surgical outcomes in ovarian cancer using AI, based on the variables that are available for clinicians prior to the surgery. Such a tool could have direct clinical relevance. Diverse models that can predict the outcome of the surgery were investigated and developed. The developed AI models were also compared against the standard statistical prediction model, which demonstrated that the AI model outperformed the statistical prediction model: the prediction of all outcomes (complete or optimal or suboptimal) (AUC of AI: 0.71 and AUC of statistical model: 0.51), the prediction of complete or optimal cytoreduction versus suboptimal cytoreduction (AUC of AI: 0.73 and AUC of statistical model: 0.50) and finally the prediction of complete cytoreduction versus optimal or suboptimal cytoreduction (AUC of AI: 0.79 and AUC of statistical model: 0.47). The most important variables for this prediction were identified as: FIGO stage, tumour grade and histology.

The application of transcriptomic analysis to cancer research raises the question of which genes are significantly involved in a particular cancer and which genes can accurately predict survival outcomes in a given cancer. Therefore, AI techniques were employed to identify the most important genes for the prediction of Homologous Recombination (HR), an important DNA repair pathway in ovarian cancer, identifying *LIG1* and *POLD3* as novel prognostic biomarkers. Finally, AI models were used to predict the HR status for any given patient (AUC: 0.87).

This project has demonstrated that AI may have an important role in ovarian cancer. AI systems may provide tools to help clinicians and research in ovarian cancer and may allow more informed decisions resulting in better management of this cancer.

Acknowledgements

This thesis would not have been possible without the assistance and support of many great people. I would like to express my deep gratitude to my supervisor Dr. Richard Edmondson for his enthusiastic guidance in this research. Many thanks go to Dr. Joe Faith for his help and guidance.

I am also grateful to all the staff at the Northern Institute for Cancer Research for helping and supporting in any form. I also thank to all my friends and colleagues who have supported me and helped me.

I give my deepest love and appreciation to my mother, and my family for their unconditional love and support. Finally, I am deeply indebted to my dear partner Juila for her love, patience and encouragement all these years.

Abbreviations

AI	Artificial Intelligence
ANN	Artificial Neural Network
AUC	Area Under Curve
BN	Bayesian Network
CT	Computed tomography
DNA	Deoxyribonucleic Acid
DT	Decision tree
FIGO	International Federation of Gynecology and Obstetrics
FN	False negative
FP	False positive
FS	Feature Selection
GA	Genetic algorithm
HR	Homologous recombination
IG	Information Gain
KNN	K- Nearest Neighbour
LOOCV	Leave One Out Cross Validation
MAE	Mean Absolute Error
MDT	Multi-disciplinary team
ML	Machine learning
MRI	Ultrasound and Magnetic Resonance Imaging
OS	Overall survival
OV_TMA_2000	Newcastle ovarian cancer tissue micro array 2000
OV_TMA_2005	Newcastle ovarian cancer tissue micro array 2005

PARP Poly (ADP-ribose) polymerase

PC Principal component

PCA Principal Component Analysis

RMSE Root Mean Squared Error

RNN Recurrent Neural Network

ROC Receiver Operating Characteristic

RPP Relative Projection Pursuit

TN True negative

TP True positive

TPP Targeted Projection Pursuit

Table of Contents

Declaration	ii
Abstract	iii
Acknowledgements	iv
Abbreviations	v
Table of Contents	vii
List of Figures	x
List of Tables	xiii
List of Equations	xv
Chapter 1 Introduction	1
1.1 Introduction	2
1.2 Ovarian cancer	2
1.2.1 Current understanding	2
1.2.1.1 Stage	2
1.2.1.2 Grade	5
1.2.1.3 CA125 Biomarker	6
1.2.1.4 Tumour Histology	7
1.2.1.5 Radiography	7
1.2.1.6 Age	8
1.2.1.7 Physical examination	9
1.2.1.8 Surgery	10
1.2.1.9 Other possible contributors	12
1.2.1.10 Standard management	12
1.2.1.11 Outcomes of the management	13
1.2.2 Clinical decision making	14
1.2.3 Future paths	15
1.3 Computing approaches	16
1.3.1 Machine learning (ML)	17
1.3.1.1 Nearest Neighbour approach (Decision tree)	18
1.3.1.2 Neural networks	25
1.3.1.3 Bayesian Networks	32
1.3.2 Feature Selection	37
1.3.2.1 Principal Component Analysis (PCA)	38
1.3.2.2 Information Gain (IG)	44
1.3.2.3 Targeted Projection Pursuit (TPP)	47
1.3.2.4 Genetic algorithm (GA)	51
1.4 Research outline	58
1.4.1 Statement of the problem	58
1.4.1.1 Complexity of the ovarian cancer management	58
1.4.1.2 Prediction of survival	58
1.4.1.3 Surgical prediction	59
1.4.2 Research questions	59
1.4.3 Structure of the thesis	59
Chapter 2 Data and Data handling	61
2.1 Introduction	62

2.2 Datasets	62
2.2.1 Newcastle ovarian cancer tissue micro array 2000 (OV_TMA_2000).....	62
2.2.2 Newcastle ovarian cancer tissue micro array 2005 (OV_TMA_2005).....	67
2.3 Data pre-processing.....	72
2.3.1 Handling missing values	72
2.3.2 Outliers.....	72
2.3.3 Coding the input values.....	74
2.3.4. Normalization and scaling.....	74
2.4 Comparison of datasets	75
2.5 Summary of chapter	78
Chapter 3 Experimental methodology	79
3.1 Introduction.....	80
3.2 Estimates of prediction performance	80
3.2.1 Resubstitution Validation.....	80
3.2.2 Hold-out validation	81
3.2.3 K-Fold Cross-Validation.....	82
3.2.4 Leave-One-Out Cross-Validation (LOOCV).....	83
3.2.5 Repeated k-fold Cross-Validation.....	84
3.2.6 Evaluation of the estimates	84
3.3 Prediction performance measurements	85
3.3.1 Possible outcomes of a model.....	85
3.3.2 Accuracy	86
3.3.3 Recall.....	87
3.3.4 Precision.....	87
3.3.5 Mean Absolute Error (MAE)	88
3.3.6 Root Mean Squared Error (RMSE).....	88
3.3.7 Receiver Operating Characteristic (ROC)	89
3.4 Over fitting	90
3.5 Process definition.....	92
3.6 Summary of the chapter	94
Chapter 4 Prognostic Data	95
4.1 Introduction.....	96
4.2 Finding the best feature set	96
4.2.1 The optimum feature set size	96
4.2.2 The most important features in the dataset	99
4.2.2.1 One feature sets	99
4.2.2.2 Two features sets.....	101
4.2.2.3 Three features sets.....	103
4.2.2.4 Four features sets.....	105
4.2.2.5 Five features sets	107
4.2.3 Discovering best features using feature selection techniques.....	109
4.3 Prediction of survival for many categories	110
4.3.1 Analysis of 64 survival categories	110
4.3.2 Analysis of Overall Survival (OS)	115
4.4 Categorization of the continues data	117
4.5 Model analysis by increasing the number of data.....	119
4.6 The best model	120
4.7 Comparison with conventional statistics.....	123
4.8 Comparison with pervious work	125
4.9 Summary of the chapter	125
Chapter 5 Surgical Prediction	128
5.1 Introduction.....	129

5.2 Statement of the problem	129
5.3 Residual prediction.....	131
5.3.1 Residual prediction: Complete/Optimal/Sub-optimal cytoreduction.....	132
5.3.2 Residual prediction: Complete and Optimal versus Suboptimal cytoreduction	139
5.3.3 Residual prediction: Complete versus Optimal or Sub-optimal cytoreduction	145
5.4 Discovering important factors.....	151
5.5 Summary of the chapter	151
Chapter 6 Gene analysis for classification of Homologous recombination	153
6.1 Introduction	154
6.2 statement of the problem.....	154
6.3 Dataset.....	156
6.4 Most important Genes	160
6.5 HR prediction models	162
6.6 Summary of the chapter	165
Chapter 7 Conclusion and Future Works.....	166
7.1 Introduction	167
7.2 Conclusions of the Thesis	167
7.3 Summary of main contributions.....	167
7.3.1 The review of related problems	167
7.3.2 The solution for prediction.....	168
7.3.3 The prototype implementation and evaluation.....	169
7.4 Further work.....	170
7.5 The overall achievement of the thesis.....	171
Appendices	173
Appendix 1 Prototype snapshot	173
Appendix 2 class diagrams	174
References	178

List of Figures

Figure 1-1 Age associated incidence of ovarian cancer	9
Figure 1-2 KNN method example (artificial dataset)	20
Figure 1-3 Decision Tree example	24
Figure 1-4 Biological neuron	27
Figure 1-5 Artificial model of a neuron	27
Figure 1-6 Feed Forward ANN model	29
Figure 1-7 Neural network for AND function	31
Figure 1-8 Connections in Bayesian network	35
Figure 1-9 Bayesian network of probability of raining today and tomorrow	36
Figure 1-10 Transformation of dataset using PCA	39
Figure 1-11 The TPP process	48
Figure 1-12 original view of SRBCT dataset using TPP	50
Figure 1-13 separated class view of SRBCT dataset using TPP	50
Figure 1-14 Significant variables in SRBCT discovered using TPP	50
Figure 1-15 GA process	55
Figure 1-16 GA process example	56
Figure 2-1 The issue of outliers in the dataset	73
Figure 2-2 Two datasets comparison	77
Figure 2-3 Survival vs. Outcome of surgery 2000-2005	77
Figure 3-1 Resubstitution Validation	81
Figure 3-2 Hold-out Validation	82
Figure 3-3 4-fold Cross-Validation	83
Figure 3-4 Leave-One-Out Cross-Validation	84
Figure 3-5 The possible outcomes of prediction	86
Figure 3-6 Receiver operating characteristic (ROC)	89
Figure 3-7 Stop training time	92
Figure 3-8 Process definition	93
Figure 4-1 the optimum feature set size (Accuracy)	98
Figure 4-2 the optimum feature set size (MAE)	98
Figure 4-3 the optimum feature set size (RMSE)	98
Figure 4-4 one feature sets performance results	100
Figure 4-5 two features sets performance results	102
Figure 4-6 three features sets performance results	104
Figure 4-7 four features sets performance results	106

Figure 4-8 five features sets performance results	108
Figure 4-9 Prediction performance results for 64 categories of survival.....	112
Figure 4-10 ROC curves: 2 survival categories	114
Figure 4-11 ROC curves: 4 survival categories	114
Figure 4-12 OS prediction performance: ROC curves.....	116
Figure 4-13 Effects of increasing number of cases on accuracy of the model	120
Figure 4-14 Prediction performance comparison (ANN, DT and BN).....	122
Figure 4-15 Model comparison using ROC curve	122
Figure 4-16 ANN vs. Logistic Regression.....	124
Figure 4-17 ANN vs. Logistic Regression (ROC curves).....	124
Figure 5-1 ROC curve (ANN): Complete / Optimal /Sub-optimal cytoreduction (CA125 not categorized).....	135
Figure 5-2 ROC Curve (Logistic Regression): Complete / Optimal /Suboptimal cytoreduction (CA125 not categorized).....	135
Figure 5-3 ROC curve (ANN): Complete / Optimal /Suboptimal cytoreduction (CA125 categorized).....	138
Figure 5-4 ROC Curve (Logistic Regression): Complete / Optimal /Suboptimal cytoreduction (CA125 categorized)	138
Figure 5-5 ROC curve (ANN): (Complete or Optimal) cytoreduction versus Sub-optimal cytoreduction (CA125 not categorized).....	141
Figure 5-6 ROC curve (Logistic Regression): (Complete or Optimal) cytoreduction versus Sub-optimal cytoreduction (CA125 not categorized).....	141
Figure 5-7 ROC curve (ANN): (Complete or Optimal) cytoreduction versus Sub-optimal cytoreduction (CA125 categorized).....	144
Figure 5-8 ROC curve (Logistic Regression): (Complete or Optimal) cytoreduction versus Sub-optimal cytoreduction (CA125 not categorized).....	144
Figure 5-9 ROC curve (ANN): Complete cytoreduction versus (Optimal or Sub-optimal) cytoreduction (CA125 not categorized)	147
Figure 5-10 ROC curve (Logistic Regression): Complete cytoreduction versus (Optimal or Sub-optimal) cytoreduction (CA125 not categorized).....	147
Figure 5-11 ROC curve (ANN): Complete cytoreduction versus (Optimal or Sub-optimal) cytoreduction (CA125 categorized).....	150
Figure 5-12 ROC curve (Logistic Regression): Complete cytoreduction versus (Optimal or Sub-optimal) cytoreduction (CA125 categorized)	150
Figure 6-1 Selective cytotoxicity of PARP inhibitors	155

Figure 6-2 First view of dataset using TPP	161
Figure 6-3 Separated class view of the dataset using TPP.....	161
Figure 6-4 HR prediction ROC curves using ANN	162
Figure 6-5 HR prediction ROC curves using BN	163
Figure 6-6 HR prediction ROC curves using DT	164

List of Tables

Table 1-1 International Federation of Gynecology and Obstetrics (FIGO) staging system for Ovarian cancer.....	4
Table 1-2 Five year survival rate for different stages of Ovarian cancer.....	5
Table 1-3 Grading of ovarian cancer	6
Table 1-4 Radiographic characteristics of Adnexel masses.....	8
Table 1-5 Characteristics of Pelvic mass on physical examination	10
Table 1-6 Outcomes of the surgery	12
Table 1-7 Outcomes of the management	14
Table 1-8 Play Golf dataset.....	22
Table 1-9 AND function truth table	31
Table 1-10 Joint probability results using Bayesian theory	36
Table 1-11 Artificial dataset for explaining PCA process	40
Table 1-12 Normalized artificial dataset for explaining PCA process	40
Table 1-13 The amount of information required to assign objects to each class for variable ‘Outlook’	45
Table 1-14 Artificial dataset as an example for GA process	55
Table 2-1 OV_TMA_2000.....	63
Table 2-2 OV_TMA_2000 Age	64
Table 2-3 OV_TMA_2000 FIGO stage	64
Table 2-4 OV_TMA_2000 Grade	65
Table 2-5 OV_TMA_2000 Histological subtype.....	65
Table 2-6 OV_TMA_2000 CA125	66
Table 2-7 OV_TMA_2000 Outcome of the surgery.....	66
Table 2-8 Analysed data for collection of OV_TMA_2005 dataset	67
Table 2-9 OV_TMA_2005.....	68
Table 2-10 OV_TMA_2005 Age	69
Table 2-11 OV_TMA_2005 FIGO stage	69
Table 2-12 OV_TMA_2005 Grade	70
Table 2-13 OV_TMA_2005 Histological subtype.....	70
Table 2-14 OV_TMA_2005 CA125	71
Table 2-15 OV_TMA_2005 Outcome of the surgery.....	71
Table 3-1 AUC results interpretation	90
Table 4-1 Important features using feature selection techniques.....	109
Table 4-2 Prediction performance results for 64 survival categories (Accuracy)	112

Table 4-3 OS prediction performance.....	116
Table 4-4 Prediction performance: Categorized data versus Continuous data	118
Table 5-1 Prediction performance of results for complete/optimal/sub-optimal cytoreduction when CA125 is not categorized using ANN	134
Table 5-2 Prediction performance results for complete/optimal/suboptimal cytoreduction when CA125 is not categorized using Logistic Regression.....	134
Table 5-3 Prediction performance results for complete/optimal/sub-optimal cytoreduction when CA125 is categorized using ANN	137
Table 5-4 Prediction performance results for complete/optimal/sub-optimal cytoreduction when CA125 is categorized using Logistic Regression	137
Table 5-5 Prediction performance results for (complete or optimal) cytoreduction versus sub-optimal cytoreduction when CA125 is not categorized using ANN.....	140
Table 5-6 Prediction performance results for (complete or optimal) cytoreduction versus sub-optimal cytoreduction when CA125 is not categorized using Logistic Regression	140
Table 5-7 Prediction performance results for (complete or optimal) cytoreduction versus sub-optimal cytoreduction when CA125 is categorized using ANN.....	143
Table 5-8 Prediction performance results for (complete or optimal) cytoreduction versus suboptimal cytoreduction when CA125 is categorized using Logistic Regression	143
Table 5-9 Prediction performance results for complete cytoreduction versus (optimal or sub-optimal) cytoreduction when CA125 is not categorized using ANN.....	146
Table 5-10 Prediction performance results for complete cytoreduction versus (optimal or sub-optimal) cytoreduction when CA125 is not categorized using Logistic Regression.....	146
Table 5-11 Prediction performance results for complete cytoreduction versus (optimal or sub-optimal) cytoreduction when CA125 is categorized using ANN	149
Table 5-12 Prediction performance results for complete cytoreduction versus (optimal or sub-optimal) cytoreduction when CA125 is categorized using Logistic Regression	149
Table 5-13 The most important variable identified using feature selection methods ...	151
Table 6-1 Gene analysis dataset.....	159
Table 6-2 12 most differentially regulated genes identified between 4 HR competent and 4 HR deficient tumours using three methods.....	161
Table 6-3 ANN prediction performance results (HR).....	162
Table 6-4 BN prediction performance results (HR).....	163
Table 6-5 DT prediction performance results (HR).....	164

List of Equations

Equation 1-1 Entropy using the frequency table of one variable.....	21
Equation 1-2 Entropy using frequency table of two variables.....	21
Equation 1-3 Information Gain of variable X.....	21
Equation 1-4 The value of the net (artificial neuron).....	26
Equation 1-5 The output of the net	26
Equation 1-6 Sum of Squared Error (SSE)	28
Equation 1-7 Calculating the error of the neuron.....	29
Equation 1-8 Adjusting the weights.....	30
Equation 1-9 Bayesian theory	33
Equation 1-10 Covariance of X and Y	41
Equation 1-11 Covariance matrix	41
Equation 1-12 Amount of information required to assign an object to class P or N	44
Equation 1-13 the entropy needed to classify objects in S_i subsets.....	44
Equation 1-14 Information gained by dividing on attribute A.....	45
Equation 1-15 definition of projection p in TPP	49
Equation 1-16 Euclidean norm.....	49
Equation 2-1 zero-mean normalization.....	75
Equation 2-2 max-min normalization	75
Equation 3-1 True error rate of k-fold cross-validation.....	82
Equation 3-2 True Negative Rate.....	86
Equation 3-3 False Negative Rate.....	86
Equation 3-4 Accuracy.....	87
Equation 3-5 Recall of given class a.....	87
Equation 3-6 Precision of given class a	88
Equation 3-7 Mean Absolute Error	88
Equation 3-8 Root Mean Squared Error	89
Equation 3-9 Area under ROC curve (AUC).....	90

Chapter 1 Introduction

1.1 Introduction

This chapter of the thesis introduces ovarian cancer. The disease is briefly described along with a discussion of the current understanding of the disease. Furthermore, the standard management of this cancer is introduced and the process of clinical decision making process is discussed. Finally, current challenges in ovarian cancer are highlighted.

The second section of this chapter then discusses artificial intelligence, specifically, machine learning and feature selection techniques. The relative merits of different methods are outlined.

Finally this chapter draws these two areas into a series of research questions.

1.2 Ovarian cancer

According to Cancer Research UK (2008a), 190,000 new cases of ovarian cancer are reported each year worldwide; representing about 4% of all cancers diagnosed in women. In the past 30 years there has been an 18% increase in ovarian cancer patients reported in the UK (Cancer Research UK, 2008a). Ovarian cancer is the 5th most common cancer in females in UK, with 6,800 cases each year, the UK has one of the highest incident rates in Europe (Cancer Research UK, 2008b). Although there has been a modest improvement in ovarian cancer treatments in the past 20 years, the long term survival rates have changed very little. In the UK only 36% of the patients will live for at least 5 years after diagnosis (Cancer Research UK, 2008c), although this rate varies for different stages of the cancer and the survival rate is around 90% in early stages compared to 15% at the late stages (Ghaemmghami, Orton and Soutter, 2003). The cost of ovarian cancer treatment in the UK and other European countries is unclear; however as an example, the Australian government estimates a lifetime treatment cost per case of \$19,677 in 2000-01 (AIHW, 2006).

1.2.1 Current understanding

This section of the thesis investigates the current understanding of ovarian cancer. The different characteristics of the cancer and the standard management of this cancer are included in the following sections.

1.2.1.1 Stage

Ovarian cancer has been divided into four categories based on anatomical distribution of the disease, called stages by the 'International Federation of Gynecology and Obstetrics' (FIGO) (Chi and Hoskins, 2000). These stages are expressed in Roman numerals from

stage I (the least advanced stage) to stage IV (the most advanced stage). Stage is a useful classification as it dictates not only prognosis but is used to determine treatment Leitao and Barakat (2009).

Table 1-1 summarizes the characteristics of different stages of ovarian cancer (Chi and Hoskins, 2000).

Stage	Characteristics
I	Growth limited to the ovaries
A	Growth limited to one ovary; no ascites; no tumour on the external surface; capsule intact
B	Growth limited to both ovaries; no ascites; no tumour on the external surfaces; capsule intact.
C	Tumour either Stage I A or I B. but with tumour on the external surface of one or both ovaries; or with capsule ruptured; or with malignant cells in ascites or peritoneal washings.
II	Growth involving one or both ovaries with pelvic extension.
A	Extension and/or metastases to the uterus and/or tubes
B	Extension to other pelvic tissues
C	Tumour either stage II A or II B, but with tumour on the surface of one or both ovaries; or capsule ruptured; or with malignant cells in ascites or peritoneal washings.
III	Tumour involving one or both ovaries with peritoneal implants outside the pelvis and/or positive retroperitoneal or inguinal lymph nodes. Superficial liver metastasis equals stage III. Tumour is limited to the true pelvis but with histologically verified malignant extension small bowel or omentum.
A	Tumour grossly limited to the true pelvis but with histologically confirmed microscopic of abdominal peritoneal surfaces
B	Tumour involving one or both ovaries with histologically confirmed implants of abdominal peritoneal surfaces, none exceeding 2 cm in diameter. Nodes are negative.
C	Abdominal implants greater than 2 cm in diameter and/or positive retroperitoneal or inguinal nodes.
IV	Growth involving one or both ovaries with distant metastases. If pleural effusion is present there must be positive cytology to allot a case to stage IV. Parenchymal liver metastases equals stage IV.

Table 1-1 International Federation of Gynecology and Obstetrics (FIGO) staging system for Ovarian cancer

(Chi and Hoskins, 2000)

The majority of cases present at high (FIGO III/IV) stage, Simon (2005). The 5 year survival rate varies for different stages of the cancer. Patients with early stage (I) ovarian cancer have 80-90% chance of five year survival (Table 1-2).

Table 1-2 summarizes the five year survival chance for different stages of ovarian cancer (Chi and Hoskins, 2000).

Stage	Five year survival
I	80-90%
II	80%
III	15-40%
IV	5-20%

Table 1-2 Five year survival rate for different stages of Ovarian cancer
(Chi and Hoskins, 2000)

1.2.1.2 Grade

Grade is an important prognostic factor for ovarian cancer and refers to how abnormal the cancer cells look when viewed through a microscope (Larma and Gardner, 2006). Normal cells grow and develop into mature cells capable of undertaking their physiological role. This process is referred to as differentiation (Smith et al., 2009). Cancer cells may look similar to normal cells (or well differentiated) or underdeveloped (or poorly differentiated). Grade generally correlates with outcome, poorly differentiated tumours being associated with a worse outcome than well differentiated tumours (Smith et al., 2009). Based on ovarian cancer grade categories, the higher the grade, the more likely it is that cancer will spread (Larma and Gardner, 2006).

Table 1-3 summarizes the characteristics of different grades of ovarian cancer (Larma and Gardner, 2006).

Grade 1 (Well differentiated)	looks similar to normal ovarian tissue
Grade 2 (Moderately differentiated)	looks less like ovarian tissue
Grade 3 (Poorly differentiated)	does not look like ovarian tissue

Table 1-3 Grading of ovarian cancer
(Larma and Gardner, 2006)

1.2.1.3 CA125 Biomarker

CA125 is a serum tumour marker. Tumour markers are generally used for four main purposes, to screen for the disease, to diagnose the disease, to monitor the response to treatment, and to diagnose recurrence.

As a tumour marker CA125 has a relatively high sensitivity but a poor specificity meaning that its role as a screening and diagnostic test has limitations (Hogdall et al., 2008).

Research (Seltzer, 1999) suggested that CA125 levels are elevated in 80% of women with epithelial ovarian cancer. The normal level of CA125 is 30-35 U/mL (unit per millilitre)(Helzlsouer et al., 1993). Zurawski et al. (1988) described an increase of CA125 level prior to clinical detection of ovarian cancer.

Although overall CA125 has a high sensitivity this is not the case for early stage disease where CA125 is often unhelpful (Helzlsouer et al., 1993). Another concern is that CA125 levels are often increased slightly as a result of other conditions including ovulation, menstruation, fibroids and endometriosis (Muyldermans et al, 1995), meaning that this test is really only applicable to postmenopausal women. On the other hand, most of the ovarian cancer cases are postmenopausal women (Cancer Research UK, 2008b). As a result CA125 remains the most widely used biomarker in the management of ovarian cancer.

Zurawski et al. (1990) and Skates et al. (1995) tried to improve the sensitivity of the marker by using the levels of CA125. They concluded many women with initially

elevated CA125 levels have stable CA125 levels over time, in comparison with ovarian cancer patients where CA125 levels were steeply rising during the study.

There are other methods available for improving the sensitivity of CA125, including decreasing the standard level of CA125 to 20 U/mL (Bourne et al., 1994) or using CA125 along with other biomarkers (Zhang et al. 2004, Hogdall et al. 2008 and Su et al. 2007).

1.2.1.4 Tumour Histology

The histology of the tumour affects the treatment of ovarian cancer (Farley and Birrer, 2010). Among all ovarian cancer tumour histology, 50-60% is reported as serous. The less common histology subtypes include endometrioid (25%), mucinous (4%) and clear cell (4%) (Seidman and Kurman, 2003). Clear cell and mucinous subtypes tend to present at limited sites (one or both ovaries) (Kikkawa et al, 2006), on the other hand serous and endometrioid tend to present at advanced stages of the cancer in many sites of the abdomen and pelvis (Seidman and Kurman, 2003).

1.2.1.5 Radiography

Imaging of the pelvic mass is regarded as one of the most important diagnostic tool for investigating ovarian cancer (Simon, 2005). There are many methods that are used to image ovarian cancer such as Computed tomography (CT) scan, Ultrasound and Magnetic Resonance Imaging (MRI).

CT scan is widely used for the primary diagnosis of ovarian cancer (Williams et al., 2010). CT utilises a series of x rays taken in a 360° sweep, each of which estimates tissue density in a 1D plane. These estimates then undergo tomography using the Radon transformation to reconstruct a 2D or 3D image.

CT gives a quick and accurate measurement of disease within the abdominal cavity and provides hard images which can be used to give serial measurements to assess treatment response by comparison with post treatment images, unlike ultrasound.

The imaging of the pelvis and abdomen remains an important factor for staging and diagnosis of this cancer.

Table 1-4 summarizes the radiographic characteristics of Adnexal masses (Simon, 2005). This table compares the size, consistency, septation and other characteristics of benign and malignant masses.

	Benign	Malignant
Size	<8 cm	>8 cm
Consistency	Cystic	Solid or cystic and solid
Septation	Unilocular	Multilocular
Bilateral/Unilateral	Unilateral	Bilateral
Other	Calcification (especially teeth)	Ascites

Table 1-4 Radiographic characteristics of Adnexal masses
(Simon, 2005)

1.2.1.6 Age

Ovarian cancer is rare among girls and young women (under the age of 30); however it increases with age (Elmasry and Gayther, 2007).

Figure 1-1 illustrates the association between age and incidence of ovarian cancer. The risk of ovarian cancer under the age of 30 is low. From the age of 30 to 50 the incidence increases slowly. Peak incidence is 60.5 per 100,000 at the age of 75 to 79.

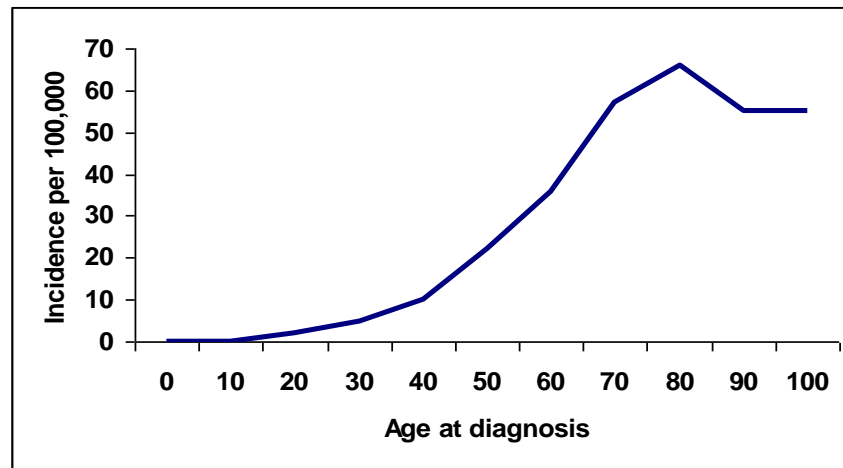


Figure 1-1 Age associated incidence of ovarian cancer
(Elmasry and Gayther, 2007)

Ovarian cancer is rare for ages under 30. For the age group of 75 to 79 it reaches the maximum value of 60.5 per 100,000.

1.2.1.7 Physical examination

Ovarian cancer has been called ‘the silent killer’ as 70% of cases are diagnosed at an advanced stage of the cancer (Edmondson and Todd, 2008). Patients may present with weight loss or vague lower abdominal pain and abdominal enlargement (Simon 2005), therefore physical examination can play an important role in first line diagnosis of the disease. Patient will be tested for mobility, consistency, bilateral and cul-de-sac involvement of the pelvic mass.

Table 1-5 **Error! Reference source not found.** summarizes the characteristics of Pelvic mass on physical examination (Simon, 2005). This table compares these characteristics for benign and malignant masses.

	Benign	Malignant
Mobility	Mobile	Fixed
Consistency	Cystic	Solid or firm
Bilateral/Unilateral	Unilateral	Bilateral
Cul-de-sac	Smooth	Nodular

Table 1-5 Characteristics of Pelvic mass on physical examination
(Simon, 2005)

1.2.1.8 Surgery

The detection of the pelvic mass as discussed in previous sections (1.2.5 and 1.2.7) is relatively easy, however characterization of that mass as benign or malignant has remained a challenge. Based on the author's observation in the operating theatre and literature (Edmondson and Todd, 2008; Balega and Shepherd, 2007; Sood and Gershenson, 2005), surgery is the most important part of the detection and diagnosis process which leads to an appropriate multimodal treatment approach. According to Sood and Gershenson (2005) "Although surgery alone is curative in only a small

number of patients, it remains the most important modality of treatment in an individual with a suspicious pelvic mass found to be ovarian cancer”.

The proper staging of the disease which was discussed in the previous section (1.2.1.) is performed at surgery. In many cases surgery results in upstaging the cancer as the tumour was more spread than the initial examination revealed (Sood and Gershenson, 2005).

The main goal of the surgery is to accurately define the stage of the disease and remove as much tumour as possible (Edmondson and Todd, 2008). However, as discussed previously (1.2.7) most patients present at stages of the cancer and as a result tumour is spread to other sites. Therefore the complete removal (complete cytoreduction) is not achievable. Complete cytoreduction is achieved when there is no macroscopic residual tumour left during surgery. The other possible outcomes of surgery are optimal cytoreduction (residual disease less than 1-2cm in diameter) and sub-optimal cytoreduction (residual disease more than 1-2cm in diameter) (Edmondson and Todd, 2008).

Table 1-6 summarizes the association between the residual disease and outcomes of the surgery. The role of the surgery will be discussed in more details in chapter five of this thesis.

At this stage, it is important to distinguish between two key words that play an important role in this investigation: Prognostic and Predictive Factors. According to Cianfranco and Goldstein (2004) prognostic factors are any measurements available before surgery that determines the overall survival rate before any treatment or therapy. In contrast, any measurement that calculates the response to a therapy is called predictive factor.

Outcome of the surgery	Residual disease
Complete cytoreduction	no macroscopic residual tumour
Optimal cytoreduction	less than 1-2cm in diameter
Sub-optimal cytoreduction	more than 1-2cm in diameter

Table 1-6 Outcomes of the surgery

1.2.1.9 Other possible contributors

There are other possible risk factors to be considered in investigations, diagnosis and determining the survival rate in ovarian cancer, such as: combination of biomarkers (Zhang et al., 2004, Hogdall et al., 2008; Su et al., 2007), race, diet, viral and familial factors (Chu and Rubin, 2005). These risk factors can be employed in the proposed model if appropriate.

1.2.1.10 Standard management

The first line management of ovarian cancer is surgery. Surgery as discussed previously is the main factor in determining the stage of the disease. According to Simon (2005), “patients with early stages of the cancer (stage one and two) with well or moderately differentiated tumours require no additional therapy after surgery”. Surgery is carried out to remove as much of the tumour as possible and may include removing one or both ovaries and if the disease is in its advance stages, removing fallopian tubes and uterus. In most advanced cases a second line surgery is required (Thackery, 2005).

After surgery, chemotherapy will be used to target cells that have spread to other organs. Chemotherapy is given usually in a day centre, one day every 21 days for six cycles. Blood tests are taken to ensure bone marrow recovery before each cycle (Le et al., 2006). Chemotherapy is usually employed after surgery when bowel function has recovered. There are no advantages in delaying chemotherapy because of wound healing (Rothenberg et al., 2003). Research (Rothenberg et al., 2003) also suggested that there are no further advantages in continuing the chemotherapy after six cycles. During the chemotherapy process, bone marrow function and the level of CA125 will be investigated.

Vasey et al. (2004) suggested that chemotherapy should continue for a maximum of six cycles if the levels of CA125 continue to fall. Following surgery chemotherapy is the recommended treatment for ovarian cancer, however in research (Rodriguez, 1992) it is suggested that in some cases radiation therapy has been useful.

1.2.1.11 Outcomes of the management

There is a need to measure the success and failure of a treatment for a cancer patient. Maurizi et al. (1999) suggested the possible measurable outcomes of the disease management.

Table 1-7 describes the possible outcomes of the management (Maurizi et al., 1999). Defining these outcomes is important as the success of a treatment is measured using these measurable end points.

Outcomes	Description
Tumour response rate	The number of patients whose tumours responded to treatment.
Death from disease	Number of patients that died because of the disease.
Disease free survival	Period of time that no cancer was detected after specific treatment.
Progression free survival	Period of time that disease stopped spreading.
Overall survival rate	The patients who are alive for a certain period of time after treatment.

Table 1-7 Outcomes of the management
(Maurizi *et al.*, 1999)

1.2.2 Clinical decision making

Thompson and Dowding (2002) defined clinical decision making as choosing between alternatives. There are different types of models available in the clinical decision making literature; The Information Processing Model (Joseph and Patel, 1990), The Intuitive-Humanist Model (Benner, 1982; Benner, 1984; Young, 1987) and O’Neill’s Clinical Decision-Making Model (O’Neill *et al.*, 2005). The investigation of these techniques is out of the scope of this research, however the first model (The Information Processing Model) is used in this research as this method is the best fitted to the purpose of this research. This model is widely used for clinical decision making (Banning, 2008).

This model uses decision trees to numerically assess the potential outcome of each decision or situation. For each tree the potential outcome will be calculated and based on that the decision will be made. Decision making in cancer needs a specialist team to evaluate and suggest possible treatments for the disease, this team is called an MDT (multi-disciplinary team). MDT members may vary depending on the case, and may include surgeons, clinical and medical oncologists, Radiologists, Pathologists, Palliative care specialists and MDT coordinator and secretarial support (Blazeby et al., 2006).

For ovarian cancer treatment, the similar MDT reviews the referral patients and based on the markers (that have been discussed previously) the possible treatment for the patient will be suggested (surgery, chemotherapy or no action). After surgery each case is also discussed at MDT meetings and possible further treatments are considered for each patient.

1.2.3 Future paths

In previous sections (1.2.1 and 1.2.2) the current understanding and standard management of ovarian cancer were discussed. This investigation highlights some possible issues in the field of ovarian cancer. Some of these issues are as follows:

- As the number of markers, treatments and the sequences of treatments are rising (such as new chemotherapy agents), it is getting more complex for the human brain to perform clinical decision making. There is the potential need for an expert computer system (eg. Artificial Intelligence) capable of investigating the possible outcome for each marker, treatment and sequence of the treatment (Friedman, 2009).
- The prediction of overall survival of an individual patient is a very difficult task. There may be some benefits in predicting the overall survival as this prediction may change the path of treatment as well as inform the patient about likely prognosis. Therefore a system that can predict the overall survival of a patient is required (Teramukai et al., 2007; Chi et al., 2008; Gerestein et al., 2009).
- As discussed previously (section 1.2.1.8), there are three different outcomes of surgery. The goal of the surgery is to remove as much tumour as possible. Therefore there are many benefits mentioned or suggested for a system that can predict the outcome of the surgery (Allen (2010); Leitao and Barakat, 2009).

1.3 Computing approaches

Current improvements in biological sciences research, especially in the related fields of genetics, genomics; transcriptomics and proteomics are producing large amounts of data. Bioinformatics is the science of developing and applying computational systems and methods for investigating and analyzing this data (Baldi and Brunak, 2001) and also it can be defined as “The science of informatics as applied to biological research” (Bioinformatics, 2008).

Artificial Intelligence (AI) introduces many powerful methods for solving important problems in bioinformatics. AI techniques can help researchers to achieve the goal of extracting useful information from the available data by building high performance and suitable models.

According to Narayanan, et al. (2002), most bioinformatics problems are seeking better solutions than currently available rather than finding the absolutely best or correct solution. AI techniques provide the methods to fit the nature of bioinformatics problems and try to find better solutions.

As discussed in previous sections, there is a need for an expert (AI) system to evaluate, analyse and suggest the outcome of each treatment path or the importance of each marker for predicting the outcome of the treatment path. Abbod, et al. (2007) compared traditional regression statistics to AI methods in the area of cancer research and suggested that AI methods are more accurate and explorative for analysing such data. They concluded that AI methods can allow researchers to individually predict disease behaviour. AI methods are already widely used for different aspects of cancer research: Radiotherapy (Cancer Research UK, 2007), control of cancer cells (Nicolini, et al., 1989), Detection (Petricoin et al., 2002), Diagnosis (Setiono, 1996) and DNA microarray classification and analysis (Mukherjee et al., 1999, Greer and Khan, 2004). AI has also been used for different types of cancer: Oral cancer (Speight et al., 1995), Lung cancer (Zhou et al., 2002), Prostate cancer (Adam et al., 2002), Breast cancer (Pena-Reyes and Sipper, 2000) and Bladder cancer (Catto et al., 2003). AI systems are also used for different aspects of ovarian cancer such as: Early detection and serum analysis (Petricoin et al. 2002, Wulfkuhle, et al., 2003), Diagnosis (Vlahou et al., 2003) and Screening (Lewis and Menon, 2003).

AI systems differ from standard computer algorithms in that they incorporate Machine Learning. Machine learning can be defined as the following “A computer program is

said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E” (Alpaydin, 2004). Machine learning systems tend to learn from existing data and adjust themselves for new data entries; also it is possible for them to fix existing errors in their design to achieve the required outcome. As discussed in the first section, clinical decision making for ovarian cancer consists of learning from past cases together with possible new approaches to suggest the possible treatment plan for the patient. It is understandable that clinical decision making perfectly fits into the machine learning concept.

1.3.1 Machine learning (ML)

Machine learning methods are the computer programs that improve their performance according to experiences (Alpaydin, 2004). The strengths of machine learning are as follows (Alpaydin, 2004):

- In some cases the only way to define a task is by example. It may be possible for researchers to provide the input and output but unable to define the relation between them. In such cases machine learning can help to explore the data and define the rules and relations.
- In large datasets it is possible to miss an important relation or rule between the data. Machine learning approaches can analyse most of the relations and help to find the hidden rules.
- It is impossible for human minds to analyse the vast amount of data generated in the current research environment.
- Environments change over time. Machines that can adapt to a changing environment reduce the need for constant redesign (self-adaptive AI).

In one view ML methods can be divided into two categories: Unsupervised Learning (e.g. clustering) and Supervised Learning (e.g. classification). We may be given a set of observations with the aim of establishing the existence of classes or clusters in the data (Unsupervised Learning) and in general this is called clustering. Or we may know for certain that there are so many classes, and the aim is to establish a rule so we can classify a new observation into one of the existing classes (Supervised Learning). In other words according to Coppin (2004), in supervised learning, the goal is to learn the route from input data to an output which is correctly provided (by a supervisor). On the other hand, in unsupervised learning, the only available data is input data and the aim is

to find rules in input data to generally find out what happens and what does not (i.e. find patterns and trends in the data). The definition of these two types of learning is essential at this point. In ovarian cancer decision making involves some markers that will lead to an outcome (supervised learning) and there is uncertainty involved in prediction of the outcome of new markers or treatment paths (unsupervised learning).

There are two distinctive terms in Machine Learning: Deductive and Inductive systems. Munakata (1998) suggests that in deductive systems, experts will provide the rules and the outcome will be determined by applying those rules to the input data; in contrast, inductive systems are the systems where there is no need for any experts and rules will be discovered by the system itself. In other words in deductive systems general rules will be introduced, the system will be trained using some examples and the trained system will be put in practice. On the other hand, in inductive systems, by using examples systems discover the rules in data and the discovered rules will be put into practice. The following sections of this chapter investigate some of the well-known ML methods.

1.3.1.1 Nearest Neighbour approach (Decision tree)

This section, firstly, introduces the nearest neighbour as a classifier. Later on, using this approach, the decision tree method is discussed.

The k-nearest neighbour (KNN) mechanism is very simple. An object (e.g. a patient) is classified by the majority vote of its neighbours, with the object being assigned to the class most common amongst its k nearest neighbours. Enas and Choi (1986) described KNN algorithm as the following:

(1) Determine parameter k = number of nearest neighbours

The choice of k is the most critical aspect of this method. A small number for this variable may increase the influence of the noise on the results. On the other hand, a large value for k , increases the computation time and also violates the philosophy of this method (the close instances may belong to the same class) (Duda et al., 2000).

(2) Calculate the distance between the query-instance and all the training samples. Sort the distance and determine nearest neighbours based on the k -th minimum distance

(3) Gather the category Y of the nearest neighbours

(4) Use simple majority of the category of nearest neighbours as the prediction value of the query instance

Selection of k as an odd number helps to find the majority of the category. In this research the prediction value can be the survival rate of the patient or the outcome of the surgery.

KNN mechanism is very simple to perform and according to Efron (1983) it is effective for a large training set, however the determination of the number of nearest neighbours is hard (Enas and Choi, 1986). The computation cost is quite high as we need to compute the distance of each query instance to all training samples.

Figure 1-2 illustrates a very simple example of KNN method on an artificial dataset.

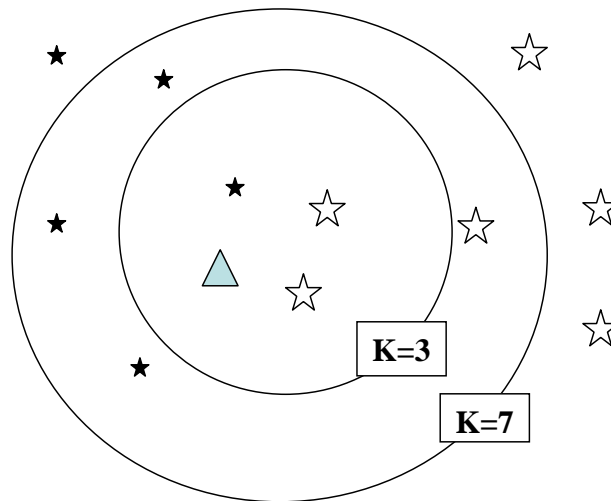


Figure 1-2 KNN method example (artificial dataset)

This example is trying to classify new data (triangle) into ‘big star’ or ‘small star’ classes. In the KNN mechanism, if we assume that the $k=3$ then new data will be classified as big star class (2 big stars and 1 small star as nearest neighbours) and if we consider $k=7$ then new data will be classified as small star class (3 big Stars and 4 small stars as nearest neighbours). This simple artificial example highlights the importance of k selection in this method.

The above approach introduces a simple KNN; however in clinical decision making there are many attributes to be considered. In order to use KNN for multiple features (attributes) in an efficient way, it is essential to introduce the term “decision trees”. In the past the decision tree approach was considered as the most successful machine learning technique in practical applications (Munakata, 1998). According to Alpaydin (2004), a decision tree is a classifier in the form of a tree, which has nodes and a root. Each node is either a leaf node (indicates the value of outcome) or decision node (which performs some tests to be carried out on a single attribute with a branch or sub-tree for each outcome). The process starts from the root node of the tree and moves through decision nodes and branches until reaching the leaf node (outcome).

Above algorithm uses “Entropy” and “Information Gain” to construct the tree (Rokach and Maimon, 2008).

$$E(S) = \sum_{i=1}^c -p_i \log_2 p_i$$

Equation 1-1 Entropy using the frequency table of one variable
(Rokach and Maimon, 2008)

S is the variable, c is the number of classes, p is the probability of the occurrence of the class number i

$$E(T, X) = \sum_{c \in X} P(c)E(c)$$

Equation 1-2 Entropy using frequency table of two variables
(Rokach and Maimon, 2008)

T is the target class, X variable, c value(s) of variable X, P(c) probability of the c, E(c) entropy of c

$$G(T, X) = E(T) - E(T, X)$$

Equation 1-3 Information Gain of variable X
(Rokach and Maimon, 2008)

T is the target class, E is the entropy

To clarify the construction of the tree processes consider the following well known dataset (Quinlan, 1986) (Table 1-8).

Outlook	Temp	Humidity	Windy	Play Golf
Rainy	Hot	High	False	No
Rainy	Hot	High	True	No
Overcast	Hot	High	False	Yes
Sunny	Mild	High	False	Yes
Sunny	Cool	Normal	False	Yes
Sunny	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes
Rainy	Mild	High	False	No
Rainy	Cool	Normal	False	Yes
Sunny	Mild	Normal	False	Yes
Rainy	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Sunny	Mild	High	True	No

Table 1-8 Play Golf dataset
(Quinlan, 1986)

In order to construct the tree following steps have to be performed:

Step 1: Calculate the Entropy of target class (Play Golf).

$$E(\text{Play Golf}) = 0.94$$

Step 2: Calculate Information Gain of each variable.

$$G(\text{Outlook}) = 0.247, G(\text{Temp}) = 0.029, G(\text{Humidity}) = 0.152, G(\text{Windy}) = 0.048$$

Step 3: Use the variable with largest value of 'Gain' as root node.

Outlook holds the largest value of 'Gain'.

Step 4: Split the branches:

4a: A branch with 'Entropy' more than 0 needs more splitting

4b: A branch with 'Entropy' 0 is a 'Leaf' node.

Step 5: Continue splitting process on none-'Leaf' branches to classify all the data.

Finally above steps leads us to the decision Tree (Figure 1-3).

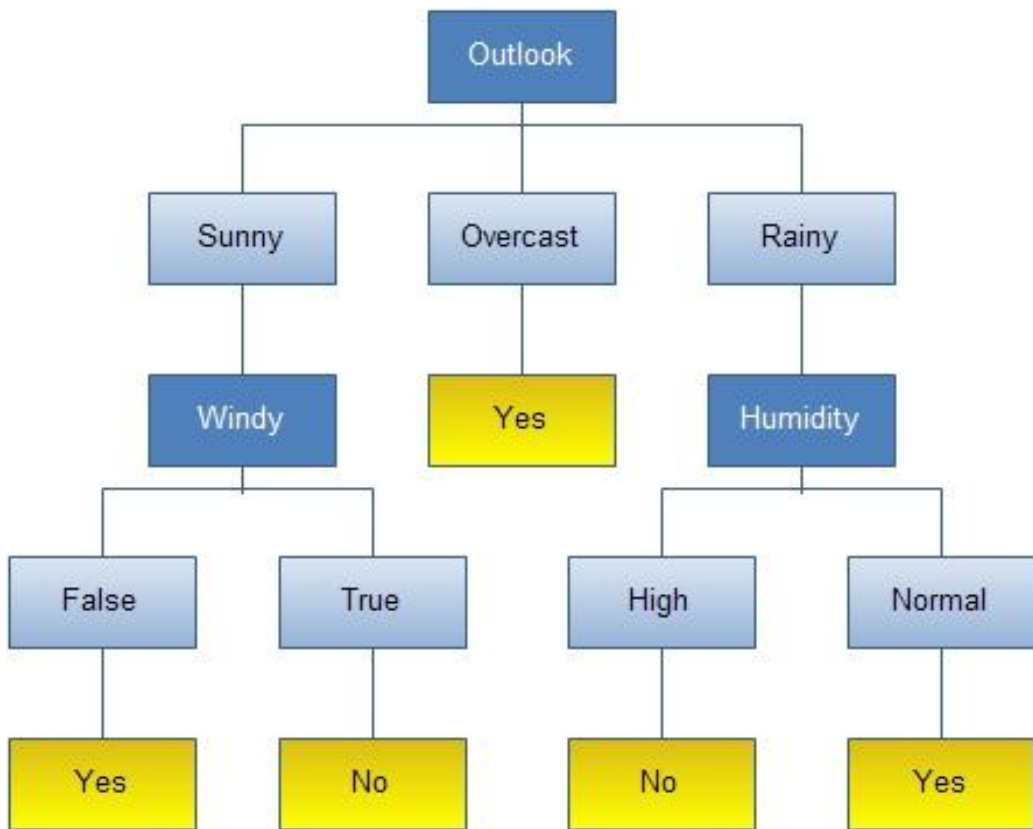


Figure 1-3 Decision Tree example
(Quinlan, 1986)

Decision trees and nearest neighbour are widely used in the field of cancer research (Adam et al.,2002; Mangasarian et al., 1995; Vlahou et al., 2003; Liu et al. , 2011).

According to Munakata (1998) decision trees are able to introduce understandable rules and identify the most important fields for prediction. Meanwhile, this method is easily understandable for every user, the important variables can be easily identified and well fitted to clinical decision making process. However according to Alpaydin (2004) decision trees can be computationally expensive to train, they are less appropriate for predicting continuous variables and are very sensitive to errors for prediction process of small data sets. Therefore, it is essential to prepare as much data as possible for this research, and making sure the continuous variables are categorized prior to prediction.

In summary, the KNN and decision trees are widely used in the field of cancer research. These methods of prediction are easy to understand for a non-professional user. This method is also well fitted to the clinical decision making process.

1.3.1.2 Neural networks

Neural networks (NN) are a very abstract computer representation of the human brain. In other words “NNs are parallel information processing structures that attempt to emulate certain performance characteristics of the biological neural system” (Naguib and Sherbet, 2001). The part of the brain considered to be the fundamental functional source of intelligence is called a neuron (Fausett, 1994), similarly neural networks consist of artificial neurons. Although the term neural network is widely used in literature of some other authors, to distinguish from the natural brain neural networks, researchers use the term Artificial Neural Networks (ANN).

The biological neuron, as the basic part of the brain, consists of three components: dendrites, soma and axon (Fausett, 1994). Dendrites are responsible for receiving signals from other neurons via axon, soma receives the inputs and when enough inputs are received and the needed operation performed, produces an output which will be sent to other neurons via output axon. Figure 1-4 illustrates a biological neuron (Fausett, 1994).

Similarly in the mathematical model, dendrites are represented by different input values x_1, x_2, \dots, x_m . Each input has a corresponding weight and the product $x_i w_i$ will be fed into the neuron (Figure 1-5).

x_1, x_2, \dots, x_m are inputs, y is output, w represents the weight, Σ the value of the net, T is the transfer function

Then the neuron calculates the net by adding up all the products for $i=1 \dots m$, so the net value can be calculated using Equation 1-4.

$$net = x_1 w_1 + \dots + x_m w_m$$

Equation 1-4 The value of the net (artificial neuron)
 x is input, w is weight

The calculation of the net is the same as the calculation of the scalar (dot) product of vectors x and w . Finally the output value ‘Out’ will be calculated (Equation 1-5).

$$Out = f(net)$$

Equation 1-5 The output of the net

This function is called a transfer (activation) function, that may be a threshold function (only passes the information), or it may be a continuous function of the combined input (Naguib and Sherbet, 2001).

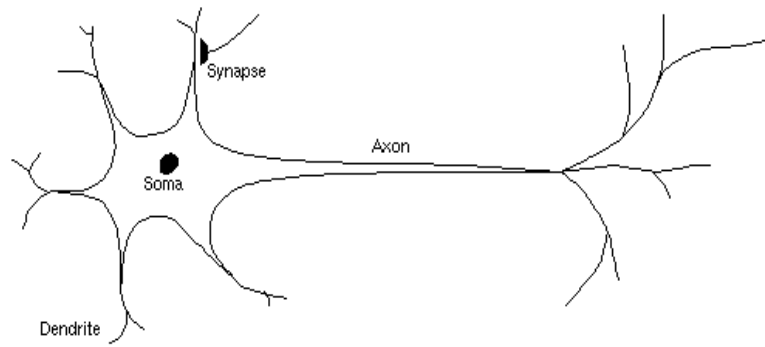


Figure 1-4 Biological neuron
(Fausett, 1994)

Dendrites are responsible for receiving the signal, soma processes the signal and axon transfer the appropriate signal to other neurons

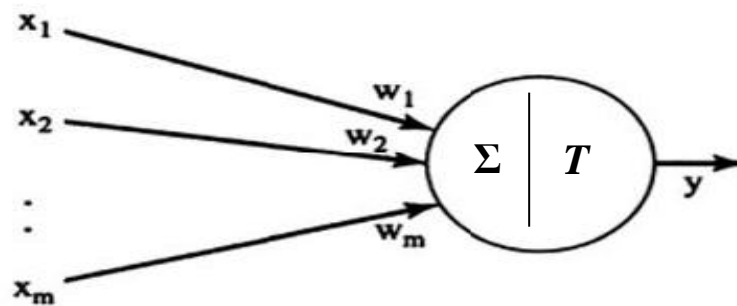


Figure 1-5 Artificial model of a neuron
(Fausett, 1994)

Σ is the value of the net (Equation 1-5) and T is the threshold of the neuron

This model of ANN is called a ‘Feed forward’ model as the process only moves in one direction from inputs to outputs. There are no connection between outputs and inputs (Lee, 2008). It contains ‘Input Layer’, ‘Output Layer’. The model may contain none, one or more than one ‘Hidden Layer(s)’. The ‘Feed forward’ ANN is the simplest and most used model of ANN (Wang et al., 2010; Lee, 2008). There are other forms of ANN such as: Recurrent Neural Network (RNN). RNN forms a ‘directed cycle’ between the connections in the network (Lee, 2008). This research uses the Feed forward ANN based on its simplicity for implementation and popularity.

Figure 1-6 illustrates a Feed forward ANN model. This model consists of one hidden layer. There are number of input neurons (N_i), hidden neurons (N_h) and output neurons (N_o) according to the problem in hand. Each input neuron is connected to all hidden neurons and respectively each hidden neurons are connected to all output neurons. All connections carry a weight that normally will be assigned as a random number in a specific range. After setting up the network is ready to learn (ready to be trained). This model of NN is a supervised learning process and according to Fausett (1994) “A neural network learns patterns by adjusting its weights.”

‘Back Propagation’ is the best known and most often used, as the learning method of multi layered feed forward ANN (Sekino and Nitta, 2007; Adali and Haykin, 2010). The goal of the learning method is to adjust the weights of the network in such a way that it minimises the ‘Sum of Squared Error (SSE)’ (Adali and Haykin, 2010). The SEE can be calculated using Equation 1-6.

$$SSE = \sum_{i=1}^n (A_i - P_i)^2$$

Equation 1-6 Sum of Squared Error (SSE)
(Adali and Haykin, 2010)

n is the number of instances for training, A is the actual value, P is the predicted value.

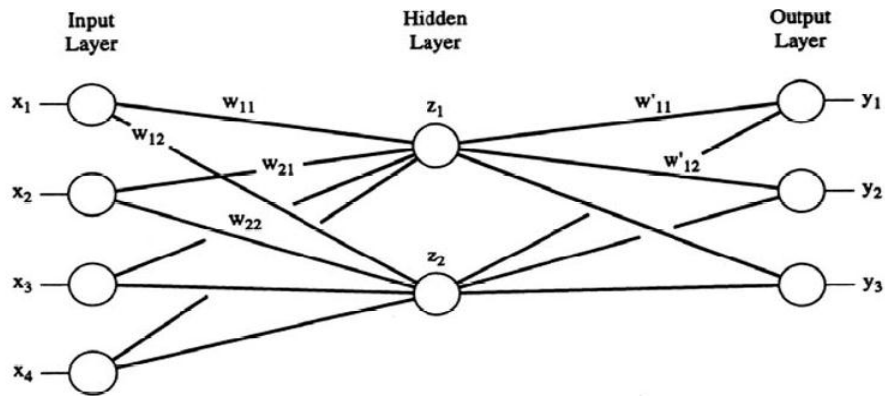


Figure 1-6 Feed Forward ANN model
(Fausett, 1994)

This model moves only in one direction (from inputs x , to outputs y).

The learning process of the network can be defined in the following steps (Adali and Haykin, 2010; Fausett, 1994):

Step 1: Randomly initial the weights of the network

Step 2: Calculate the output. Given a set of inputs (I), calculate the predicted value (P).

Step 3: Calculate the ‘error’. The error for output neurons is calculated as A-P (The difference between actual value and predicted value). The error of hidden or input neuron can be calculated using. ∴

$$B = \sum_{i=1}^n b_i * w_i$$

Equation 1-7 Calculating the error of the neuron
(Adali and Haykin, 2010)

n is the number of neuron(s) connected to a given neuron, b is the error of target neuron, w is the weight of connection between given and target neuron.

Step 4: Adjust the weights. The new weights can be calculated using Equation 1-8.

$$w_{i,j} = w_{i,j} + r * b_j * f_j(I_j) * o_i$$

Equation 1-8 Adjusting the weights

w_{i,j} is the weight of connection between neuron i, j. r is the learning rate (user defined value of how much the weights can be modified), b_j is the error of neuron j, f_j(I_j) is the value of activation function of neuron j for the input I_j, o_i is the output of neuron I during the first step.

To clarify the process, the following simple example, demonstrates the process of the above method. Consider the ‘AND’ function. This function gets two possible binary inputs (x and y) and generates an output based on the input values. Table 1-9 summarizes the all possible inputs and outputs of AND function.

The goal is to generate an ANN model to satisfy all the possible answers to AND function (Table 1-9). It is understandable that this network can be constructed using a neuron (Figure 1-7 a). After creating the network, network has to be trained to identify the values of w1 and w2 and also the threshold value of the function. There are infinite possible answers to this problem; however one possible network is illustrated in Figure 1-7 b.

x	y	output
0	0	0
0	1	0
1	0	0
1	1	1

Table 1-9 AND function truth table

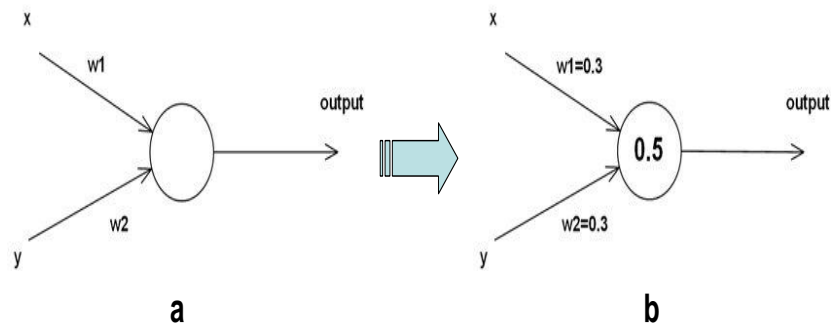


Figure 1-7 Neural network for AND function
(a) the initial generated model, (b) trained and ready to use network

Let's say that $x=1$ and $y=0$ as input values. The output values have to be 0 as '1 AND 0=0'. So we have $1 \times 0.3 + 0 \times 0.3 = 0.3$ which is less than the 0.5 threshold value so the output is 0. For $x=1$ and $y=1$ the value 0.6 is greater than 0.5 so the output is 1. The network produces the correct outcome for all possible inputs so it is trained and ready to predict new values.

ANN is widely used in different fields of cancer research and for many different cancers: Wilding et al. (1994), Fogel et al. (1995), Zhou et al. (1997), Maclin and

Dempsey (1991), Maclin and Dempsey (1992), Burke et al. (2001) Djavan et al. (2002), Kawakami et al. (2008), Stephan et al. (2008); Hoyt et al. (2010) and Ayer et al., 2010.

Similar to all methods NN has its strengths and weaknesses. NN are robust and can work if the training data contain errors (Burke et al., 2001) and also in most cases, the prediction accuracy is high (Fausett, 1994). NN provides a great power to predict the outcome on an individual basis (Naguib and Sherbet, 2001). According to McLaren et al. (2009) NN can capture non-linear relationships in the dataset.

On the other hand, NN training time can be considerably high (Burke et al., 2001) and there is no guarantee that the network is trainable (Fausett, 1994). According to Naguib and Sherbet (2001) in most cases it is hard to understand the weights used in the network.

In summary, NN as a prediction model is widely used in the field of cancer research. The prediction accuracy of this model is high and also it can handle errors in the dataset. The generated model can be adopted in such a way that it can handle new treatments and biomarkers. Considering that the most important goal for this research is prediction accuracy, the training time is not an issue and there is not necessity for the user to understand the weights of the network.

1.3.1.3 Bayesian Networks

“A Bayesian network is a graphical model for probabilistic relationships among a set of variables” (Neapolitan, 2004). There is a massive improvement in all branches of science, however in many cases there are still some uncertainties (Baldi and Brunak, 2001). For example in cancer research the researchers are uncertain about the outcome of the new treatment or influences of a new biomarker. Bayesian networks can be the solution to many of these problems for the following reasons:

Bayesian networks can handle incomplete data (Neapolitan, 2004). Most of the modelling approaches can effectively model and investigate the input data if a complete dataset is provided, however if one of the main input objects is not provided those models can not sufficiently model the data. On the other hand, when using Bayesian network it is possible to solve the problem. This advantage is very important especially in medical research as the input or outcome of some variables may not be complete at the start of the research.

“Bayesian networks allow one to learn about causal relationships” (Jensen, 2001).

Understanding the casual relation will help us to fully understand the problem domain as well as making predictions about the activities even if they didn't perform yet.

Bayesian networks provide facilities that make the modelling of the prior knowledge straightforward (Baldi and Brunak, 2001). The modelling of the current knowledge about the data is facilitated in this method. This makes it a very powerful method as the prior knowledge about the data in connection with the relations that will be discovered using this method can fully model, understand and analyse the data.

Before describing the structure of the Bayesian network it is essential to describe Bayesian approach to probability. According to Leonard and Hsu (1999) “Bayesian probability of an event is the person's degree of belief in that event”, in contrast the classical probability of an event which is the physical property of that event. For example, classical probability considers the probability that a coin will land heads and Bayesian probability is interested in the degree of our belief that the coin will land heads. In other words, in Bayesian probability, “a probability will always depend on the state of the knowledge of the person who provides the probability” (Gelman et al., 1995). For example in a normal situation a person predicts that the probability of a coin showing heads on the next toss is 0.5, however if we convince that person that the coin is weighted in favour of heads, he would assign the a higher probability to the event. There are some criticisms and solutions to those problems of the Bayesian approach to probability which is outside the scope of this thesis.

First we are going to introduce the Bayesian theory. According to Neapolitan (2004), for two events E and F (the probability of E and F are not equal to zero) Bayesian theory is the probability of F given E, multiply by the ‘prior’ probability of E, divided by the ‘prior’ probability of F (Equation 1-9). Prior probability of E means the probability that E happens regardless of any other event.

$$P(E | F) = \frac{P(F | E)P(E)}{P(F)}$$

Equation 1-9 Bayesian theory

P(F | E) the probability of F given E, P(E) and P(F) are prior probability of events E and F.

In general according to Neapolitan (2004) a Bayesian network for a set of variables $X = \{X_1, \dots, X_n\}$ consists of a network structure S that interprets a set of conditions about variables in X and a set of P of probabilities associated with each variables. These sets together define the joint probability distribution of X . Bolstad (2004) describes the joint probability distribution as: The function $f(x,y)$ which gives the probability for a given probability of x and y within the range of values of x, y . In Bayesian networks variables are depicted as nodes, connections represent the probabilistic dependence between variables and conditional probabilities encode the strength of the dependencies. To construct a Bayesian Network we simply draw connections for a given set of variables from the cause variables to their effects and finally we determine the local probability distributions. In other words according to Neapolitan (2004) the algorithm is as following:

- 1. Choose an ordering of variables X_1, \dots, X_n
- 2. For $i = 1$ to n
 - add X_i to the network
 - select parents from X_1, \dots, X_{i-1} such that

$$P(X_i | Parents(X_i)) = P(X_i | X_1, \dots, X_{i-1})$$

It is possible to have three types of connections in the network (Koski and Noble, 2009):

- **Linear connection:** The two end variables are usually dependent on each other. The middle variable renders them independent (Figure 1-8 a).
- **Converging connection:** The two end variables are usually independent on each other. The middle variable renders them dependent (Figure 1-8 b).
- **Divergent connection:** The two end variables are usually dependent on each other. The middle variable renders them independent (Figure 1-8 c).

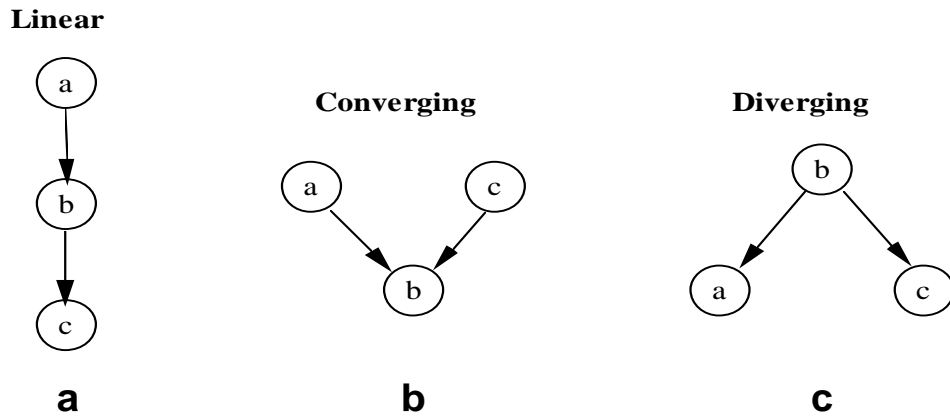


Figure 1-8 Connections in Bayesian network
a: linear connection, b: converging connection, c: diverging connection

In order to clarify the process consider a well know example of “Given a situation where it might rain today, and might rain tomorrow, what is the probability that it will rain on both days?” (Niedermayer, 1998).

It is possible to assume that raining on two consecutive days are not independent as if it is raining today it is more likely that it will rain tomorrow. Let’s assume that $P(\text{rain today})=0.20$ and $P(\text{rain tomorrow given that it rains today})=0.70$. We are going to calculate all the joint probabilities using the Bayesian theory (Equation 1-9). Table 1-10 summarizes these probabilities.

Suppose that $P(A)$ is the probability of raining today and $P(!A)$ the probability of not raining today. Similarly $P(B)$ is the probability of raining tomorrow and $P(!B)$ the probability of not raining tomorrow. Therefore, a Bayesian network can be constructed.

	Raining tomorrow	Not Raining Tomorrow	Marginal (Rain today)
Raining today	0.14	0.06	0.20
Not Raining today	0.16	0.64	0.80
Marginal(Rain Tomorrow)	0.30	0.70	

Table 1-10 Joint probability results using Bayesian theory

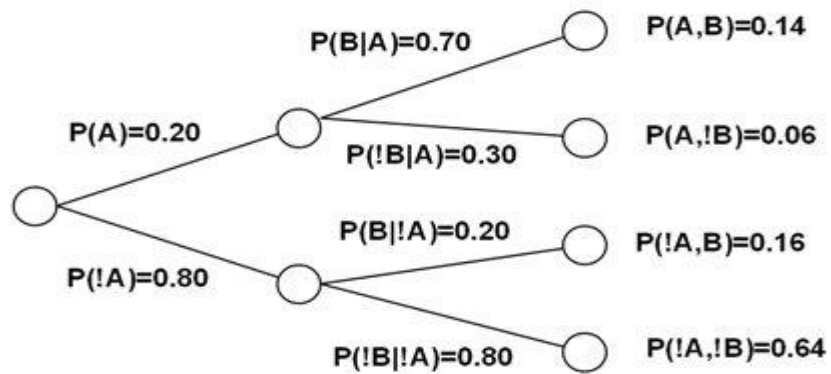


Figure 1-9 Bayesian network of probability of raining today and tomorrow
P(A) probability of raining today, *P(!A)* probability of not raining today

P(B) probability of raining tomorrow, *P(!B)* probability of not raining tomorrow

As mentioned earlier, because of the nature of this approach, many cancer researchers used this method for different aspects of cancer research such as: Kahn et al. (1997), Bulashevskaya et al. (2004), Wang et al. (1999), Jansen et al. (2003), Lacave and Diez (2003), Van der Gaag et al. (2002), Burnside, et al. (2004), Gevaert et al. (2006).

At the start of the section the potential benefits of this method have been mentioned however the initial knowledge of the probabilities (prior probabilities) are very important in this method and in some cases it is not possible to completely predict such probabilities (Niedermayer, 1998). It is common to miss the probability of an important event. It is possible that an event which plays an important role unanticipated in the network (Neapolitan, 2004). Also the potential computing cost is high (Neapolitan, 2004).

In summary Bayesian network provides a powerful realistic prediction method for predicting the outcome. Many researchers applied this method to cancer research studies. The other benefit that makes this method a powerful prediction method is that it can handle incomplete data. Therefore, based on the strong facilities that this method provides and the fact that number of researchers applied this method to cancer research, this method was selected to include in for this research.

1.3.2 Feature Selection

The following sections of the report investigate some of the well-known feature selection techniques in two categories: Multivariate and Univariate Techniques.

The next sections of the chapter investigate two most common and powerful Univariate feature selection techniques: Principal Component Analysis (PCA) and Information Gain (IG). The reason that these two techniques have been selected is that they are the most effective methods for univariate feature selection (Kantardzic (2002), Beer et al. (2002), Yeung, et al. (2001), Gabrilovich and Markovitch (2004), Changjing and Shen (2005), Cho and Ryu (2002), Ben-Dor et al. (2000), Blum and Langley (1997), Brazma and Vilo (2000) and Jensen and Shen (2004).

Later on this section investigates the multivariate techniques. Saeys et al. (2007) described multivariate techniques as: methods that consider subsets of variables (features) together and assess the dataset based on the subset of selected features. Two

techniques have been selected for investigation in this research: Targeted Projection Pursuit (TPP) and Genetic algorithm (GA). TPP is proven to be one of the most effective feature selection techniques (Enshaie and Faith, 2009) and literature review revealed that GA is one of the high performance feature selection techniques. For each technique the algorithm is discussed by using mathematical demonstrations of the technique and the process is clarified by using an example. The limitations of each technique have been discussed and applications to gene expression data have been identified.

1.3.2.1 Principal Component Analysis (PCA)

The most popular method for dimensionality reduction of a large dataset is Principal Component Analysis (PCA) (Kantardzic, 2002). The idea of PCA is to reduce the dimensionality of a data set which includes a large number of interrelated variables, while collecting as much variation as possible that is present in the data set. PCA transforms original datasets into a new set of variables called the principal components (PCs). These components are uncorrelated, and they are sorted so that the first few keep most of the variation present in all of the original variables (Jolliffe, 2002). In other words PCA explains the correlation structure of a set of interpreter variables using a smaller set of linear combinations of these variables. These linear combinations are called components (Larose, 2006).

Figure 1-10 (Scholz, 2006) is the transformation of a gene expression dataset using PCA. This technique reduces a large number of variables to a lower number of new variables known as principal components (PCs). Three-dimensional gene expression samples are projected onto a two dimensional component space that maintains the largest variance in the data. These two-dimensional visualisations of the samples allow us to make qualitative conclusions about the separation of the classes in the dataset.

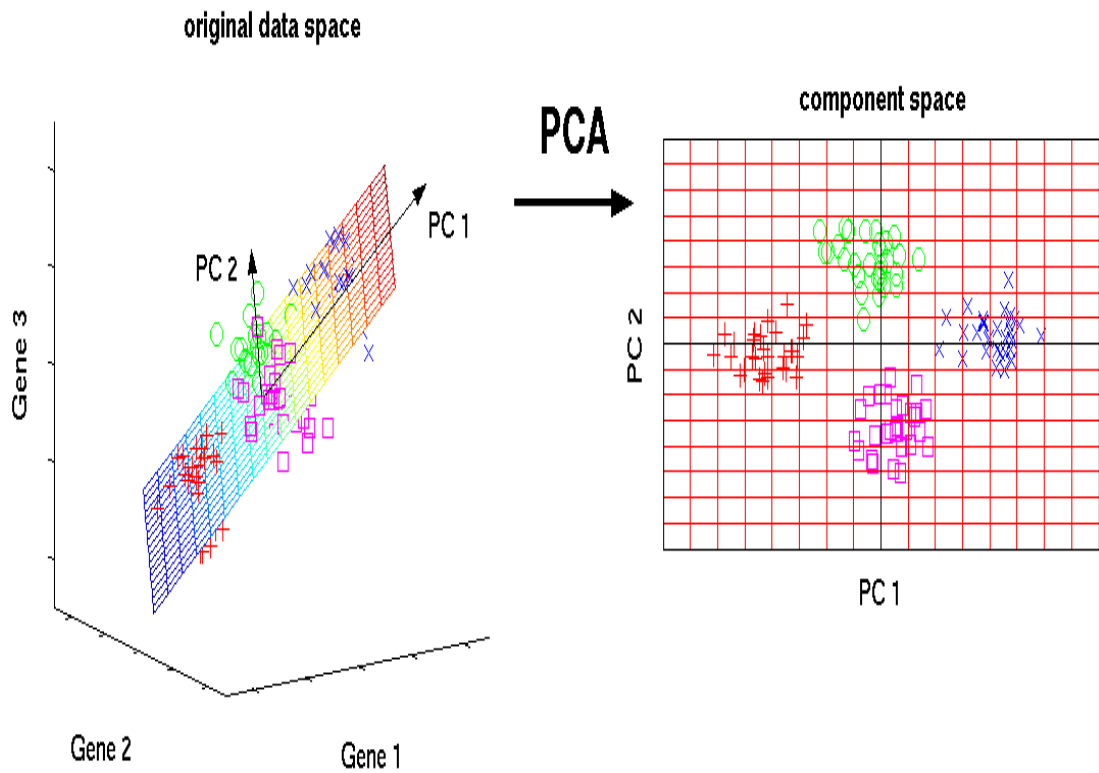


Figure 1-10 Transformation of dataset using PCA
Scholz (2006)

Chart on the left is a representation of a 3 dimensional gene expression dataset. Using PCA three dimensional dataset is transformed to a two dimensional dataset, which uses two principal component (PC1, PC2), to be plotted.

Process and example

Jolliffe (2002) introduced an informal process for applying PCA on a dataset; this section follows these steps and applies them to an artificial dataset (Table 1-11). This dataset consists of only two dimensions which made it possible for the author to demonstrate and visualize each step.

Step 1: Data

As mentioned above the dataset that has been used in this section is an own-made dataset (Table 1-11) which contains two variables.

X	2.5	0.5	2.2	1.9	3.1	2.3	2	1	1.5	1.1
Y	2.4	0.7	2.9	2.2	3	2.7	1.6	1.1	1.6	0.9

Table 1-11 Artificial dataset for explaining PCA process

Step 2: Normalization

For PCA to work properly, we have to normalize the dataset by subtracting the mean from each of the data dimensions. The mean subtracted is the average across each dimension, therefore all the x values have \bar{X} (the mean of the x values of all the data points) subtracted from them, and all the y values have \bar{Y} subtracted from them. This produces a data set that its mean is zero. This procedure is known as “zero mean” (Kantardzic, 2002). Table 1-12, demonstrates the original dataset (Table 1-11) after normalization.

X	.69	-1.31	.39	.09	1.29	.49	.19	-.81	-.31	-.71
Y	.49	-1.21	.99	.29	1.09	.79	-.31	-.81	-.31	-1.01

Table 1-12 Normalized artificial dataset for explaining PCA process

Step 3: Calculate the covariance matrix

Covariance provides a measure of the strength of the correlation between two or more sets of random variates (Spiegel, 1992). The covariance for two random variates X and Y, each with sample size N, is defined in Equation 1-10.

$$\text{cov}(X, Y) = \sum_{i=1}^N \frac{(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Equation 1-10 Covariance of X and Y

N is the sample size of X and Y, \bar{x} is the mean value of X, \bar{y} is the mean value of Y.

If we have a data set with more than 2 dimensions, there is more than one covariance measurement that can be calculated. For example, from a 3 dimensional data set (dimensions x, y, z) we could calculate cov(x, y), cov(x, z) and cov(y, z). In fact, for an n-dimensional data set, you can calculate $\frac{n!}{(n-2)! * 2}$ different covariance values (Spiegel, 1992). A useful way to get all the possible covariance values between all the different dimensions is to calculate them all and put them in a matrix (Spiegel, 1992) (Equation 1-11).

$$C^{m \times n} = (c_{i,j} = \text{cov}(\text{Dim}_i, \text{Dim}_j)) \quad C = \begin{pmatrix} \text{cov}(x, x) & \text{cov}(x, y) & \text{cov}(x, z) \\ \text{cov}(y, x) & \text{cov}(y, y) & \text{cov}(y, z) \\ \text{cov}(z, x) & \text{cov}(z, y) & \text{cov}(z, z) \end{pmatrix}$$

Equation 1-11 Covariance matrix

The equation on left represents a covariance matrix for the sample size m and equation on the right is a generated covariance matrix for sample size 3.

Keeping these in mind (Equation 1-11) we calculate the covariance matrix for our dataset:

$$\text{cov} = \begin{pmatrix} .61 & .61 \\ .61 & .71 \end{pmatrix}$$

Step 4: Calculate the eigenvectors and eigenvalues of the covariance matrix

Let A be a $n \times n$ matrix. The λ is an eigenvalue of A if there exists a non-zero vector \mathbf{v} such that $A\mathbf{v} = \lambda\mathbf{v}$. In this case, vector \mathbf{v} is called an eigenvector of A corresponding to λ . For each eigenvalue λ , the set of all vectors \mathbf{v} satisfying $A\mathbf{v} = \lambda\mathbf{v}$ is called the eigenspace of A corresponding to λ (Press et al., 1992)

The method of calculating these values is described by Press et al. (1992) and after applying those calculations on our covariance matrix we have the following:

$$\text{eigenvalues} = \begin{pmatrix} .49 \\ 1.28 \end{pmatrix}$$

$$\text{eigenvectors} = \begin{pmatrix} -.73 & -.67 \\ .67 & -.73 \end{pmatrix}$$

Step 5: Choosing components and select features

In general, once eigenvectors are found from the covariance matrix, the next step is to order them by eigenvalue, highest to lowest. This gives us the components in order of significance (Jolliffe, 2002). At this stage we can decide to ignore the components of lesser significance. By doing this we may lose some information, but if the eigenvalues are small the lost information is not significant. If we leave out some components, the final data set will have fewer dimensions than the original. Later we can pick the largest component(s) as chosen features.

PCA relies upon on some assumptions which result in some limitations for this powerful technique. These assumptions are:

- **Linearity assumption:** In PCA we assumed the observed dataset to be linear combinations of certain basis (Jolliffe, 2002). Some new methods such as kernel-PCA (Scholkopf et al., 1998) overcome this limitation by considering non-linearity of the data.
- **Importance of mean and covariance assumption:** As suggested above, PCA is based on the importance of the means of the (or) a dataset and accordingly the covariance of the values, which in some recent situations may mislead the method (Jackson, 2003).
- Sotoca, et al. (2007) investigated PCA for classification and observed that limitations of PCA arises from the fact that, in general, a principal component (for

the whole sample set) does not have to provide a high deviation rate and so therefore a good classification ability.

- Bishop (1996) also discussed some major limitations for PCA such as: PCA need a large number of samples to perform well. In PCA we have to determine how many factors to retain and in some cases it is possible to get negative values for reduced data by PCA which in some cases (such as chemical components) it cannot clearly be connected to the sources.
- The main problem mentioned (Banks et al., 2004) is that PCA can detect the linear structures in the data; however it is difficult to detect non-linear structures with PCA.

PCA as one of the most powerful feature selection techniques is widely used to analyse and process gene expression data. Bicciato, et al. (2002) applied PCA on gene expression data and their results demonstrate that the employed procedure allows the identification of specific phenotype markers and can classify previously unseen instances in the presence of multiple classes of cancer. Beer et al. (2002) used PCA to classify lung cancer data and they concluded that this technique successfully classifies the data.

However Yeung, et al. (2001) used PCA for clustering the ovarian cancer patient's data and they recommended against using PCA to reduce dimensionality of the data before applying clustering algorithms because they observed that the quality of clustering results on the data after PCA is not necessarily higher than that on of the original data, sometimes lower, unless external information is available.

PCA is used by cancer researchers to analyse and investigate the data for different aspects of cancer research: Bicciato et al. (2002), Beer et al. (2002), Skala et al. (2007), Rodríguez-Piñero et al. (2007), Yao et al. (2008), Kamath and Mahato (2009) and Hsu et al. (2009)

In summary, PCA is widely used by many researchers for solving dimensionality reduction problems and in most cases this technique is very successful. Most researchers indicate that this technique is one of the most effective techniques in feature selection.

1.3.2.2 Information Gain (IG)

Information gain is one of the most efficient measures of feature ranking in feature selection (Gabrilovich and Markovitch, 2004). IG estimates feature weights (importance of the features) determining for each feature how much information it contributes to the knowledge of the classes of the training data items.

Yang and Pedersen (1997) compared five different feature selection scores on two datasets and showed that Information Gain is amongst the two most effective. The method which we use to assign weights to the features is called Gain Ratio, a normalized variant of information gain (Daelemans, et al., 1999).

Let P and N be two classes and S a dataset with p -elements and n -elements. The amount of information needed to decide if a given example data belongs to P or N can be defined using Equation 1-12 (Larose, 2006).

$$I(p, n) = -\frac{p}{p+n} \log_2 \frac{p}{p+n} - \frac{n}{p+n} \log_2 \frac{n}{p+n}$$

Equation 1-12 Amount of information required to assign an object to class P or N
(Larose, 2006)

P and N are two classes of dataset S, with p-elements and n-elements

Let sets $\{S_1, S_2, \dots, S_v\}$ form a partition of the set S, when using the attribute A and each S_i contain p_i examples of P and n_i examples of N. The entropy (the expected information) needed to classify objects in all the subsets S_i is defined by Equation 1-13 (Larose, 2006).

$$E(A) = \sum_{i=1}^v \frac{p_i + n_i}{p+n} I(p_i, n_i)$$

Equation 1-13 the entropy needed to classify objects in S_i subsets
(Larose, 2006)

v number of subsets, p_i and n_i the examples of p and n in S_i

Therefore, the information that would be gained by dividing on attribute A can be calculated using Equation 1-14.

$$Gain(A) = I(p, n) - E(A)$$

Equation 1-14 Information gained by dividing on attribute A
I(p, n) is calculated using Equation 1-12, E(A) is calculated using Equation 1-13

Process and example

To clarify the IG process, the example (Quinlan, 1993) that was previously used (section 1.3.1.1) will be discussed. The dataset (Table 1-8) consists of four variables and an outcome class.

Using Equation 1-12, the amount of information required to assign objects to each class for variable ‘Outlook’ is calculated. Table 1-13, summarizes the results of this calculation.

Outlook	p_i	n_i	$I(p_i, n_i)$
Sunny	2	3	0.971
Overcast	4	0	0
Rain	3	2	0.971

Table 1-13 The amount of information required to assign objects to each class for variable ‘Outlook’

Therefore, using Equation 1-13, Entropy of variable 'Outlook' can be calculated as the following:

$$E(\text{outlook}) = \frac{5}{14} I(2,3) + \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) = 0.694$$

The information gain can be calculated as the following: (Equation 1-14)

$$\text{Gain}(\text{outlook}) = I(9,5) - E(\text{outlook}) = 0.246$$

Similarly:

$$\text{Gain}(\text{temperature}) = 0.029 \text{ and } \text{Gain}(\text{humidity}) = 0.151 \text{ and } \text{Gain}(\text{windy}) = 0.048$$

According to the IG mechanism attribute 'Outlook' holds most of the information by itself and therefore after sorting attributes in order of importance we have:

Attributes selection using IG: 1) Outlook 2) Humidity 3) Windy 4) Temperature

The following two problems with IG's will have a negative influence on the classification accuracy, in particular when there are many features available.

In IG if features are dependent, this will generally not be reflected in their weights (Hintz, 1991). For example a feature that contains some information about the classification class on its own, but none when another more informative feature is present will receive a non-zero weight. Features which contain little information about the classification class will receive a small weight but a large number of them may still overrule more important features (Chang et al., 2005).

IG is considered to be one of the most effective algorithms for feature selection on gene expression data. Changjing and Shen (2005) concluded that in general, IG attribute selection is beneficial for improving the performance of the common learning algorithms on gene expression data. Cho and Ryu (2002) used IG for feature section and classification of gene expression data of cancer and observed that IG is a powerful and successful method for these types of feature selections. Ben-Dor et al. (2000), Blum and Langley (1997), Brazma and Vilo (2000) and Jensen and Shen (2004) also investigated

IG mechanism in their research and observed that this method successfully selected the appropriate features.

IG is also used by researchers' (Andrew et al., 2006; Ashraf et al., 2010) to identify the most important factors for outcome prediction in cancer research.

IG is widely used for feature selection and research indicates the high performance of this technique. In gene expression data analysis literature, IG is treated as one of the most effective approaches for feature selection.

1.3.2.3 Targeted Projection Pursuit (TPP)

One of the major problems with linear methods such as PCA is that they are not able to detect non-linear structures in data. To overcome this problem projection pursuit method has been introduced which is widely viewed as an effective method which can find interesting structures, both linear and non-linear, in complex data (Banks et al., 2004). In projection pursuit, it is assumed that interesting structure is different from normal distribution (All normal distributions are symmetric and have bell-shaped density curves with a single peak (Gaten, 2000)). The reason that normal distribution is uninteresting is that for most multi-dimensional data sets, almost all low dimensional projections are normal (Diaconis and Freeman, 1984). However this assumption is not always valid as the purpose of the research determines the importance of the features (Banks et al., 2004). Sometimes we would like to find interesting features in a dataset as compared to another dataset, and it may be possible that our later dataset does not follow a normal distribution (Hyvarinen, 2001).

Normal projection searches the space of all possible projections to find out which maximises an index that measures the quality of each resulting view (Lee et al, 2005). In other words according to Friedman and Tukey (1974) projection pursuit searches for projection vectors that maximize projection indices which highlight deviations from the normal distribution.

There are some alternative approaches developed for normal projection pursuit such as: Relative Projection Pursuit (RPP) (Hiro, et al., 2002) and Targeted Projection Pursuit (TPP) (Faith, et al., 2006). The investigation of RPP is out of scope of this research however TPP method is going to be investigated in this research.

According to Faith, et al. (2006) “Targeted projection pursuit proceeds by hypothesizing an ideal view of the data, then finding a projection that best approximates that view.” This approach could be very useful as analysing all possible views for high dimensional data is almost impossible. In this approach the appropriate target view of the data is defined as the view that clearly shows the separation between known classes in the dataset.

Figure 1-11, describes the overall process of TPP in which p is the number of features in the dataset and k is the new dimension which has to follow the rule $k < p$:

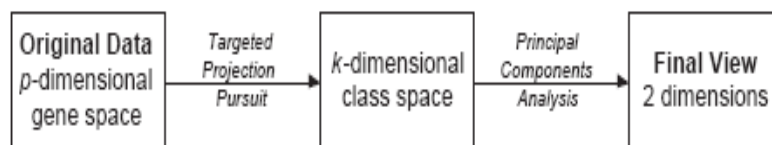


Figure 1-11 The TPP process
(Faith, Mintram and Angelova, 2006)

p is the number of variables in the original dataset

Feature Selection using TPP is achieved by finding a view that separates the class in question from all other samples, and then inspecting the resulting projection to see which genes are most significant.

Process and example

Suppose we have a $n \times f$ matrix which demonstrates the dataset S (n number of samples and f number of features in the dataset) and a target view T , a $n \times 2$ (n number of samples) matrix that describes a two dimensional view of the dataset. Therefore, a projection P ($p \times 2$ matrix) that minimizes the difference between the resulting of the projection and the desired view is required (Equation 1-15).

$$\min \|T - SP\|$$

Equation 1-15 definition of projection p in TPP
(Faith, 2007)

T the target view, *P* the projection and *S* dataset

$\| \|$ denotes Euclidean norm (Equation 1-16)

Suppose a^{ij} is the i^{th} row and j^{th} column of matrix A. Euclidean norm can be defined as Equation 1-16.

$$EN = \sqrt{\sum_{ij} a_{ij}^2}$$

Equation 1-16 Euclidean norm
(Barni et al., 2000)

A solution to Equation 1-15 may be found by “training a single layer perception with p input units and two linear output units” (Faith, 2007). In other words, for each element in the dataset, compute the corresponding row in P.

As mentioned above, TPP tends to find the views in the data. In the field of feature selection, the interesting view of the data is the view in that separation in classes in the data set is clear. By selecting this view of the dataset (view in which all classes are separated), TPP calculates and sorts the importance of each feature in the view. This process identifies the most important features and they can be used as selected features.

The process starts by loading a dataset to TPP. Figure 1-12 illustrates the original visualization of SRBCT dataset (Khan et al, 2001) (This dataset contains micro-array data of four cancers) using TPP.

After loading the dataset, by selecting the “Separate class” option, TPP introduces the view of the data in which classes in the dataset are separated. Figure 1-13 illustrates the secreted class view of SRBCT dataset using TPP. At this stage TPP calculates the significance of each feature in this view and highlights the most important features. By sorting this table, the most important features can be selected. Figure 1-14 illustrates the significance table provided by TPP. For each variable the significance of the variable was calculated.

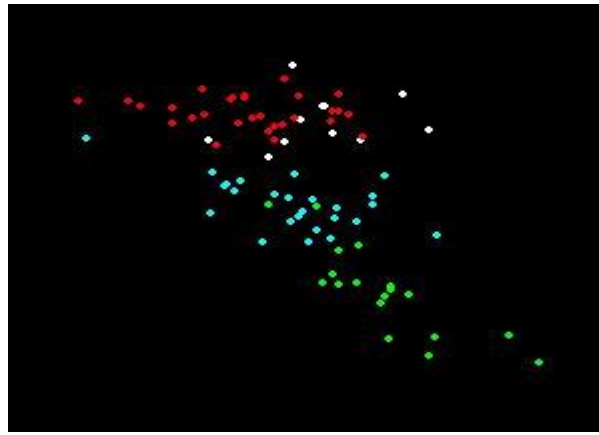


Figure 1-12 original view of SRBCT dataset using TPP

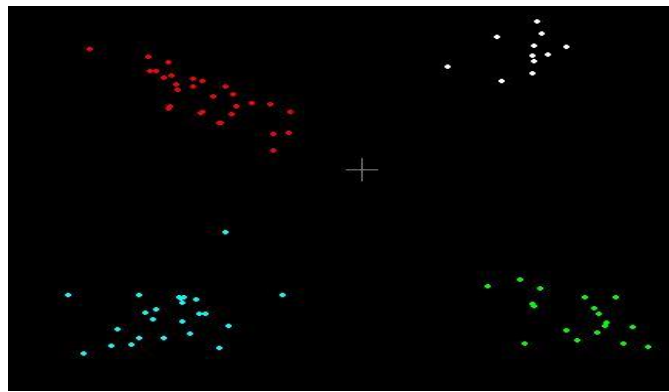


Figure 1-13 separated class view of SRBCT dataset using TPP

	Attribute	X Weight	Y Weight	Significance
11	325182 ...	0.984	-0.644	1.383
13	143586...	-0.94	0.432	1.071
39	142134 ...		-0.189	1.036
27	207274 ...	-0.51	-0.864	1.007
9	784224 ...	-0.855	-0.498	0.979
3	770394 ...	-0.841	0.47	0.928
1	812105 ...	0.802	-0.509	0.903
12	383188 ...	0.446	-0.787	0.818
47	898219 ...	-0.671	-0.578	0.784
36	866702 ...	-0.643	0.565	0.733
22	814260 ...	-0.568	0.597	0.679
26	769716 ...	-0.697	-0.415	0.658
14	377461 ...	-0.598	0.478	0.586
15	796258 ...	-0.58	-0.5	0.586
18	786084 ...	0.517	-0.544	0.563
21	296448 i...	-0.632	-0.356	0.526
20	244618 ...	-0.565	-0.429	0.503
30	295985 ...	-0.042	-0.686	0.472
46	143306 l...	-0.349	-0.526	0.399
49	208699 ...	0.597	-0.168	0.384
4	183337	0.199	0.575	0.37

Figure 1-14 Significant variables in SRBCT discovered using TPP

TPP provides a beneficial significance in case of feature selection, as this technique visualizes the dataset and makes it more understandable for researchers. Considering the benefits of visualization of the dataset, generating views for large datasets can be a very time consuming process. At this stage of the research there is no other research available to clarify the possible limitations of this technique.

Faith, et al. (2006) investigated TPP performance on 3 datasets and they are as follows:

- **LEUK:** This dataset is the result of a study of gene expression in two types of acute leukaemia: acute lymphoblastic leukaemia (ALL) and acute myeloid leukaemia (AML) (Golub et al, 1999)
- **SRBCT:** This dataset comprises cDNA microarray analysis (Khan et al, 2001).
- **NCI:** This dataset records the variation in gene expression among the 60 cell lines from the National Cancer Institute's anticancer drug screen (Scherf et al, 2000)

Results of investigations on the above datasets indicate that TPP performed better than the investigated methods.

Another research (Haddow et al., 2008) analysed the relationship between the primary structure of proteins, and their functional properties using TPP. Researchers concluded that in the case of the kinase proteins, TPP offers useful analytical results and more importantly TPP identified regions that are known to be important in differentiating functional classes.

In summary TPP techniques demonstrate a successful new method in dimensionality reduction. The views of the dataset that generated using TPP, help researchers have a better understanding of the dataset in hand. The investigated ion on the three datasets concluded a high performance for this technique. However further investigation is required to compare the performance of this technique with some other dimensionality reduction techniques in particular for feature selection.

1.3.2.4 Genetic algorithm (GA)

Genetic Algorithm (GA) was first used by John Holland and his students in 1975.

Usually Genetic Algorithms are used in optimization (finding the optimum of a function). Traditional optimization techniques like gradient descent are local in scope

(Butz et al., 2005). This means that the optima they seek are the best in a neighbourhood of the current point. This causes problems, if the function has many local optima or, even worse, if the gradient is not computable. Enumerative search methods (e.g. the A*-algorithm) on the other hand are only applicable, if the search space is discrete and not too large (Russel and Norvig, 1995). The advantages of Genetic Algorithms are that there is no need to have much information about the function to optimize and GA is applicable even if the search space is very large.

Genetic Algorithms are using random choice as a tool to guide a highly exploitative search through a discrete coding of the search space. That means the transition from one state in search space to another is probabilistic, not deterministic. According to Dechter and Mateescu (2004), deterministic matching systems use a combination of algorithms and business rules to determine when two or more records match. On the other hand, Probabilistic matching uses likelihood ratio theory to assign comparison outcomes to the correct or more likely decision. Another characteristic of Genetic Algorithms is that they don't search from a single point, but from a population of points at the same time. That means Genetic Algorithms are a global search procedure, because they explore the search space from many points in parallel. This way they can avoid local optima. However, it should be mentioned that Genetic Algorithms are not always the best choice. Depending on the problem in hand, a traditional gradient descent may actually be much faster than a computationally expensive Genetic Algorithm (Lin and He, 2005).

The first step in the implementation of genetic algorithms is to identify and model (generate) the initial population. In normal genetic algorithm each member of this population will be a binary string of length N and it will be referred to as a "chromosome" (Reeves and Rowe, 2003). The most common and easiest way to achieve this is a binary code. Consider we want to find the maximum of the objective function $K: \{0, \dots, 128\} \mapsto \mathbb{R}, x \mapsto x^2$.

Each point $x \in \{0, \dots, 128\}$ can be coded as a 7 bit binary number. By putting these points together we get a population of chromosomes as:

$$(b_1, \dots, b_7), b_i \in \{0, 1\}$$

Then each chromosome is evaluated by using a fitness function, which measures the quality of its corresponding solution. According to Reeves and Rowe (2003) the fitness function maps the outputs of the evaluation function for the different individuals in the population to the fitness value of one single individual. That means the fitness is defined with respect to other members of the population. Fitness function should always produce a positive value for the fitness and is defined as the following (Reeves and Rowe, 2003):

$$f : x \mapsto \mathbb{R}^+$$

In the normal Genetic Algorithm the fitness is defined as $\frac{\sigma_i}{\bar{\sigma}}$, where σ_i is the evaluation associated with individual and $\bar{\sigma}$ is the average evaluation of all individuals of the population (Reeves and Rowe, 2003). The fitness value can also be assigned based on the ranking of the individuals (Goldberg, 1989).

At each generation (step) the fittest (the best) chromosome will be kept, and as the process continues only the fittest chromosomes will remain (the best solution for the problem after last generation). For creating this new generation GA uses crossover and mutation operators (they are exchanging parts of their genetic information). In GA roulette wheel parent selection is used to pick parents for the new population, where each individual is represented on the wheel by a space that is proportional to its fitness. The roulette wheel is rotated and an individual is chosen until the intermediate population is filled up.

In broader usage of the term, a genetic algorithm is any population based model that uses selection and recombination operators to generate new sample points in the search space (Reeves and Rowe, 2003).

In order to clarify the GA algorithm we are going to explain the crossover and mutation operators.

Crossover: After selecting the parents we need to combine them to create new generation, this process is known as crossover. In other words crossover is a matter of replacing some of the properties in one parent by properties of the corresponding genes of the other (Reeves and Rowe, 2003). According to Reeves and Rowe (2003) there are different kinds of crossover operators used in GA such as: one point crossover, m point

crossover, non-linear crossover and generalized n point crossover. Imagine we have two strings of

$$(b_1, b_2, a_3, a_4, a_5) \text{ And } (b_1, b_2, b_3, b_4, b_5)$$

one point crossover randomly selects crossover point (in this case between 1... 5) and new solution produced by combining the pieces of the original parents as follows (crossover point=2):

$$(a_1, a_2, b_3, b_4, b_5) \text{ And } (b_1, b_2, a_3, a_4, a_5)$$

Mutation: The purpose of mutation in GA is preventing the population of chromosomes from becoming too similar to each other, thus slowing or even stopping evolution (Reeves and Rowe, 2003). Mutation simply means complementing the chosen bit(s). For example, the string 10001011 with mutation applied at genes 3 and 5, becomes 10100011.

The following is a simple GA algorithm (Jones, 1998):

1. A population of m random individuals is initialized.
2. Fitness scores are assigned to each individual.
3. Using roulette wheel parent selection m/2 pairs of parents are chosen from the current population to form a new population.
4. With probability P_c , children are formed by performing crossover on the m/2 pairs of parents. The children replace the parents in the new population.
5. With probability P_m , mutation is performed on the new population.
6. The new population becomes the current population.
7. If the termination conditions are satisfied exit, otherwise go to step 3.

Figure 1-15, illustrates the simple GA algorithm.

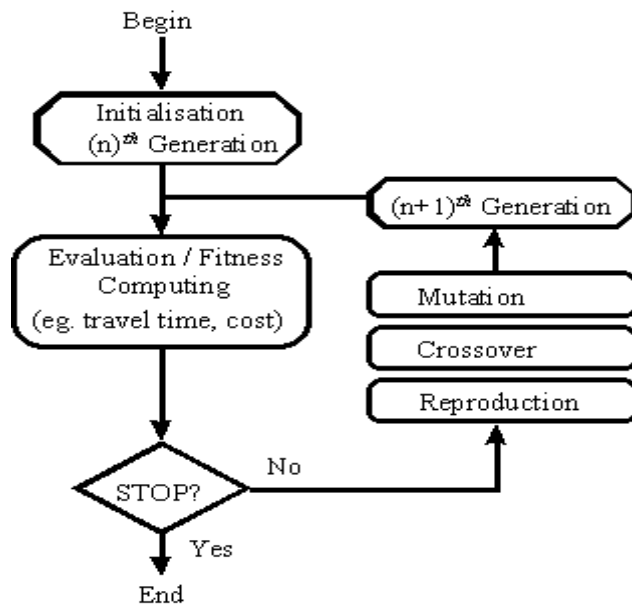


Figure 1-15 GA process
(Mitchell, 1998)

To clarify the process, consider the following artificial example (Table 1-14). The fitness values are generated based on an artificial scenario.

ID	Chromosome	Fitness
1	11000	0.80
2	00111	0.20
3	10101	0.60
4	11011	0.70

Table 1-14 Artificial dataset as an example for GA process

At this point GA will select the fittest chromosomes according to their fitness values. In this case chromosomes number one ($f(1)=0.80$), chromosomes number four ($f(4)=0.70$) and chromosomes number three ($f(3)=0.60$) will be selected.

After selection chromosomes crossover and mutate. The first chromosome gets last bit mutated and second and third chromosome crossover to produce two new ones. After these transformations population at time t_{n+1} is 11001, 10001, and 10111.

Figure 1-16 illustrates the process of the above example.

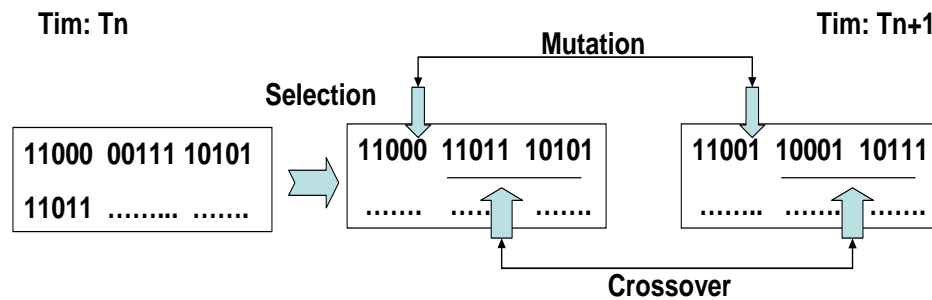


Figure 1-16 GA process example

Although genetic algorithms have proven to be an efficient and powerful problem-solving strategy, on the other hand GA has certain limitations. However, it will be shown that all of these can be overcome and none of them “bear on the validity of biological evolution” (Mitchell, 1998).

- Lack a convergence proof: It means that GA may not know when it is done (Jackson and Norgard, 2008). One possible solution to this problem is to examine some possible generations and if no improvement is observed, terminate the process.
- Premature convergence: If a chromosome that is fitter than most of its competitors emerges early in the process of the algorithm, it may drive down the population's diversity too soon, leading the algorithm to unite on the local optimum that the individual represents rather than finding the global optimum (Forrest 1993, Mitchell 1996). In other words according to Jackson and Norgard (2008) GA has the problem of “climbing local hills in the solution space”. One possible solution to this problem is to control the strength of the selection (Mitchell, 1998).
- Time and computational effort: several researchers (Holland 1992; Forrest 1993; Haupt and Haupt 2004) advise against using genetic algorithms on analytically solvable problems. This doesn't mean that GA cannot find the optimum solution for

the problem, but the traditional analytic methods take much less time and computational effort than GA.

- Suitable fitness function and operators: as mentioned in the previous section GA needs a high performance fitness function and suitable crossover and mutation operators to perform well. Failing to design and implement such properties for GA will lead algorithm to failure.

As mentioned previously (section 1.3.2) feature selection is a search problem and on large datasets such as gene expression datasets, this search is proven to be more difficult. As described in this section GA is a high performance search process which is widely applied on gene expression data sets. Chakraborty and Maka (2005) investigated GA on gene expression data and compared some other methods (such as chi squared) and concluded that GA outperformed other algorithms. Karzynski et al. (2003) claims that GA is a 100% successful method in context of diagnostic applications of DNA micro-arrays. Many other researchers(Weinberg et al. 2001, Zhu et al. 2007, Hruschka et al. 2004, Liu et al. 2004, Gesu et al. 2005, Gesu et al. 2005 and Ma et al. 2003) employed GA for feature selection or clustering the gene expression datasets and claimed that this technique is very powerful and accurate compared to other methods.

Considering efficiency and accuracy of GA, many researchers used GA on gene expression data and concluded that this technique is very effective and powerful. Although GA has some limitations, they can be resolved if they are carefully addressed during the implementation.

1.4 Research outline

The following section demonstrates the outline of the present research. The statement of the problem is introduced and the research questions are introduced. Later on, the structure of the thesis is clarified.

1.4.1 Statement of the problem

Previously (Section 1.2) the current understanding of ovarian cancer was discussed. The standard management of ovarian cancer highlights some problems and possible improvement for management of ovarian cancer. The following sections address these problems.

1.4.1.1 Complexity of the ovarian cancer management

The number of markers, treatments and the sequences of treatments are rising. Therefore it is getting more complex for human brain to perform clinical decision making. There is a need for an expert computer system (e.g. Artificial Intelligence) which is capable of investigating the possible outcome for each marker, treatment and sequence of the treatment (Friedman, 2009).

These expert systems provide a tool that may help clinicians to analyze and predict different treatment pathways. These models can act as a tool for predicting the outcome of each treatment pathway.

The other possible benefits of using such systems are that they can learn from experience and improve their prediction accuracy.

1.4.1.2 Prediction of survival

The prediction of the overall survival of a patient is a very difficult task. There may be some benefits in predicting the overall survival as this prediction may change the path of treatment. The prediction of survival may change the method of recruitment in randomised studies. Understanding the possible survival duration of the patient may also be beneficial for the patient.

Recently there were some attempts in predicting the overall survival (Gerstein et al., 2009). Although the introduced models could predict the overall survival, the accuracy of the prediction is not high (c statistic of 0.67). The introduced system is a static model that cannot improve its prediction accuracy as the number of new cases emerges.

Therefore a system that can predict the overall survival of the patient is needed (Teramukai et al., 2007; Chi et al., 2008; Gerestein et al., 2009).

1.4.1.3 Surgical prediction

As discussed previously (section 1.2.1.8), there are three different outcomes of surgery. The current management of ovarian cancer is the surgery after discovering abnormality during the physical examination, rises in CA125 biomarker and abnormality on CT scans. Only in small number of cases, due to health issues, the patient is referred for other treatment pathways rather than surgery. As discussed (section 1.2.1.8) the goal of the surgery is to remove as much tumour as possible. Based on the author's observation in theatre room and ovarian cancer literature, in some cases it is impossible to remove the entire tumour during surgery. Such patients (that complete cytoreduction surgery is not achievable) could be referred to another treatment pathway rather than primary surgery.

In the past (Bristow et al., 2000; Funt et al, 2004) there were some attempts to predict the outcome of the surgery based on CT scan results. None of these attempts could successfully introduce a reliable model. At present there is an essential need for such system that can predict the outcome of the surgery (Allen, 2010).

Therefore there are many benefits mentioned for a system that can predict the outcome of the surgery (Allen (2010); Leitao and Barakat, 2009).

1.4.2 Research questions

The research questions are derived from the problems identified in existing literature.

Thus, two of this research's questions are as follows:

1. Can Artificial intelligence and Machine learning methods accurately predict the survival rate of ovarian cancer patients?
Accuracy in this question refers to better prediction results compared to the current available models.
2. Can Artificial intelligence and Machine learning methods predict the outcome of the surgery using the variables available for clinicians prior to surgery?

1.4.3 Structure of the thesis

The thesis consists of seven chapters. **Chapter 1:** This chapter investigates the current understanding of ovarian cancer. The current management of ovarian cancer and clinical decision making regarding ovarian cancer is investigated in this chapter. The Artificial intelligence, Machine learning methods and feature selection techniques are introduced

in this chapter. Based on the identified problems in current management of ovarian cancer and the potential benefits of computational methods the research questions are identified.

Chapter 2: Introduces the dataset that will be used in this research. This chapter investigates the data pre-processing processes.

Chapter 3: Investigates the process definition of this research. This chapter also introduces the estimates of prediction performance and prediction performance measurements. In this chapter, the possible problems during the process are identified and addressed.

Chapter 4: Investigates the possibility of prediction of survival rates. The most important variables for this prediction are identified. The results of the present model are compared against the previously developed model and conventional statistical models.

Chapter 5: Investigates the prediction of the outcome of the surgery. The most important factors for this prediction are identified. The results of the prediction are also compared against the results of prediction using statistical models.

Chapter 6: In this chapter, using feature selection and Artificial intelligence models, a gene analysis for discovering most important genes for prediction of chemotherapy response is introduced.

Chapter 7: Finally this chapter summarizes the thesis and introduces the possible future work that can be completed.

Chapter 2 Data and Data handling

2.1 Introduction

This chapter of the thesis introduces the datasets that will be used in this research. Each dataset is briefly described and the characteristics of each dataset are presented. These characteristics include the number of data and the percentage of the data which is available for all the variables. The number of data for the complete dataset (the dataset with no missing values) and the entire collected data is made available.

Each dataset needed to be pre-processed prior to any analysis; therefore the pre-processing phase including the normalization of the data, handling missing values, identifying outliers and coding the input values are briefly described.

After the pre-processing phase, the datasets were merged and the initial comparison of the two datasets is described.

2.2 Datasets

To discover the hidden relations between the data, the appropriate datasets are required. The machine learning section in chapter one (section 1.3.1) describes the need for datasets to train and evaluate the created models. This section introduces the datasets that will be used to extract information from the data to build, train and evaluate the models.

2.2.1 Newcastle ovarian cancer tissue micro array 2000 (OV_TMA_2000)

The Newcastle ovarian cancer tissue micro array (OV_TMA_2000) (Wilkinson et al., 2008) is used as a primary dataset. Briefly, the OV_TMA_2000 comprises 167 cases of epithelial ovarian cancer. All tissue samples were taken from sequential patients who underwent primary surgery between 1995 and 2000 at the Northern Gynaecological Oncology Centre, Gateshead, UK, with appropriate ethical approval. All patients underwent maximal effort primary cytoreductive surgery followed by platinum based chemotherapy either with or without paclitaxel. For each tumour sample all histology reports, H&E stained slides and formalin fixed paraffin embedded blocks were retrieved and the diagnosis of epithelial ovarian cancer confirmed. 1mm cores were then taken from these areas and added to the OV_TMA_2000. Data were simultaneously extracted from the patient record to include age, FIGO stage, grade, histological subtype, preoperative CA125 and outcome of surgery. These data are available to a clinician immediately after staging surgery. The overall survival was also recorded.

41/167 cases had more than two data items missing and were therefore excluded from further analysis giving a total of 126 cases within the dataset, summarised in table 2.1.

	All available cases	Complete data	1 or 2 missing values	Useable data	Unusable data
Number of cases	167	70	56	126	41

Table 2-1 OV_TMA_2000
The available data in OV_TMA_2000 dataset

- Age: The age of the patient in years or missing. (Table 2-2)
- FIGO stage: The value is recorded as one of the following: 1a, 1b, 1c, 2a, 2b, 2c, 3a, 3b, 3c, 4 or missing. (Table 2-3)
- Grade: The possible values are: poorly differentiated, moderately differentiated, well differentiated or missing. (Table 2-4)
- Histological subtype: The possible subtypes (section 1.2.1.4) are: serous, papillary serous, mucinous, clear cell, endometrioid or missing. (Table 2-5)
- CA125: Continuous data greater than one. (Table 2-6)
- Outcome of the surgery (section 1.2.1.8): complete cytoreduction, optimal cytoreduction, sub-optimal cytoreduction or missing. (Table 2-7)

Age	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
20-30	2	1	1.20	1.43
30-40	3	1	1.80	1.43
40-50	12	6	7.19	8.57
50-60	36	13	21.56	18.57
60-70	54	27	32.34	38.57
70-80	39	16	23.35	22.86
80-90	12	5	7.19	7.14
90-100	1	0	0.60	0
100+	2	1	1.20	1.43
Missing	6	0	3.59	0
Total	167	70	100%	100%

Table 2-2 OV_TMA_2000 Age
Breakdown of available data for Age

FIGO stage	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
1a	15	4	8.98	5.71
1b	1	1	0.60	1.43
1c	11	5	6.59	7.14
2a	1	1	0.60	1.43
2b	3	1	1.80	1.43
2c	10	4	5.99	5.71
3a	6	5	3.59	7.14
3b	9	2	5.35	2.86
3c	76	40	45.51	57.14
4	18	7	10.78	10.00
Missing	17	0	10.18	0
Total	167	70	100%	100%

Table 2-3 OV_TMA_2000 FIGO stage
Breakdown of available data for FIGO stage

Grade	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
Poorly differentiated	80	39	47.90	55.71
Moderately differentiated	45	21	26.95	30.00
Well differentiated	42	10	25.15	14.29
Missing	0	0	0	0
Total	167	70	100%	100%

Table 2-4 OV_TMA_2000 Grade
Breakdown of available data for Grade

Histological subtype	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
Serous	38	15	22.75	21.43
Papillary serous	41	18	24.55	25.71
Mucinous	18	5	10.78	7.14
Clear cell	16	4	9.58	5.71
Endometrioid	54	28	32.34	40.00
Missing	0	0	0	0
Total	167	70	100%	100%

Table 2-5 OV_TMA_2000 Histological subtype
Breakdown of available data for Histological subtype

CA125	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
1-35	9	1	5.39	1.43
35-100	15	8	8.98	11.43
100-500	46	28	27.54	40.00
500-1000	25	12	14.97	17.14
1000-3000	26	12	15.57	17.14
3000+	19	9	11.38	12.86
Missing	27	0	16.17	0
Total	167	70	100%	100%

Table 2-6 OV_TMA_2000 CA125
Breakdown of available data for CA125

Outcome of the surgery	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
Complete cytoreduction	54	22	32.34	31.43
Optimal cytoreduction	30	18	17.96	25.71
Sub-optimal cytoreduction	64	30	38.32	42.86
Missing	19	0	11.38	0
Total	167	70	100%	100%

Table 2-7 OV_TMA_2000 Outcome of the surgery
Breakdown of available data for Outcome of the surgery

2.2.2 Newcastle ovarian cancer tissue micro array 2005 (OV_TMA_2005)

The original dataset (OV_TMA_2000) was discussed in the previous section (section 2.2.1). However according to Marzban (2009) and Lee (2010) availability of more data can significantly improve the performance of the models. Therefore in order to improve the models, a further dataset was required. Subsequently, with appropriate ethical approval, the patient's files which were admitted to the Northern Gynaecological Oncology Centre, Gateshead, UK, between 2000 and 2005 were analysed and appropriately recorded. The number of patient files that are available for analysis in the period of 2000-2005 is 1057. In order to capture all required data for each patient, it was essential to read and record the following for each patient:

- MDT meeting reports.
- Operation notes.
- The letters sent to each patient.

The following table (Table 2-8) summarizes the number of MDT meeting reports, operation notes and the letters sent to each patient during the 2000-2005 period that were analyzed and recorded during the collection of OV_TMA_2005 dataset. Most of this information was paper based, therefore making the analysis and the recording process very difficult and time consuming.

Number of patients	MDT reports	Patients' letters	Operation notes
1057	3300	28300	4310

Table 2-8 Analysed data for collection of OV_TMA_2005 dataset
The amount of analysed patient files for collection of the OV_TMA_2005 dataset

The data was divided into similar categories as the previous section.

	All available cases	Complete data	1 or 2 missing values	Useable data	Unusable data
Number of cases	501	248	34	282	219

Table 2-9 OV_TMA_2005
The available data in OV_TMA_2005 dataset

Similar to the OV_TMA_2000 dataset, each variable can hold different possible values:

- Age: (Table 2-10)
- FIGO stage: (Table 2-11)
- Grade: (Table 2-12)
- Histological subtype: (Table 2-13)
- CA125: (Table 2-14)
- Outcome of the surgery: (Table 2-15)

Age	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
20-30	2	0	0.40	0.00
30-40	10	3	2.00	1.21
40-50	48	20	9.58	8.06
50-60	98	42	19.56	16.94
60-70	153	79	30.54	31.85
70-80	116	54	23.15	21.77
80-90	69	46	13.77	18.55
90-100	5	4	1.00	1.61
100+	0	0	0.00	0.00
Missing	0	0	0.00	0
Total	501	248	100%	100%

Table 2-10 OV_TMA_2005 Age
Breakdown of available data for Age

FIGO stage	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
1a	68	33	13.57	13.31
1b	5	2	1.00	0.81
1c	70	31	13.97	12.50
2a	3	1	0.60	0.40
2b	2	1	0.40	0.40
2c	20	10	3.99	4.03
3a	14	9	2.79	3.63
3b	19	11	3.79	4.44
3c	230	128	45.91	51.61
4	46	22	9.18	8.87
Missing	24	0	4.79	0
Total	501	248	100%	100%

Table 2-11 OV_TMA_2005 FIGO stage
Breakdown of available data for FIGO stage

Grade	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
Poorly differentiated	225	172	47.90	44.91
Moderately differentiated	49	29	26.95	9.78
Well differentiated	63	47	25.15	12.57
Missing	164	0	0	32.73
Total	501	248	100%	100%

Table 2-12 OV_TMA_2005 Grade
Breakdown of available data for Grade

Histological subtype	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
Serous	19	19	3.79	7.66
Papillary serous	100	64	19.96	25.81
Mucinous	44	30	8.78	12.10
Clear cell	30	14	5.99	5.65
Endometrioid	206	121	41.12	48.79
Missing	102	0	20.36	0
Total	501	248	100%	100%

Table 2-13 OV_TMA_2005 Histological subtype
Breakdown of available data for Histological subtype

CA125	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
0-35	22	14	4.39	5.65
35-100	62	34	12.38	13.71
100-500	131	88	26.15	35.48
500-1000	56	31	11.18	12.50
1000-3000	71	43	14.17	17.34
3000+	46	38	9.18	15.32
Missing	113	0	22.55	0
Total	501	248	100%	100%

Table 2-14 OV_TMA_2005 CA125
Breakdown of available data for CA125

Outcome of the surgery	Number of available cases (#)		Percentage of the data (%)	
	All data	Complete data	All data	Complete data
Complete cytoreduction	207	120	41.32	48.39
Optimal cytoreduction	134	80	26.75	32.26
Sub-optimal cytoreduction	72	48	14.37	19.35
Missing	88	0	17.56	0
Total	501	248	100%	100%

Table 2-15 OV_TMA_2005 Outcome of the surgery
Breakdown of available data for Outcome of the surgery

2.3 Data pre-processing

The quality of a machine learning model or a classifier is highly dependent on many factors. Probably the most important factor is the quality and representation of the available data (Kotsiantis et al., 2006; Asyali et al., 2006). This is an important part of the research as the models such as ANN are very sensitive to the quality of the data and consequently, the predictive results can be highly affected. Therefore, the identification of undesirable data in the dataset and the removal of such data can ensure the better performance of the models. This section investigates the essential parts of data pre-processing.

2.3.1 Handling missing values

It is common for real world datasets to contain missing values (Kotsiantis et al., 2006). On some occasions due to technical issues it is impossible to calculate the value of a feature for a given instance or occasionally the data is lost based on how the data is recorded. There are many methods suggested for handling the missing values in the dataset. One of the most common methods is to remove the instances that have missing value(s) (Engelbrecht, 2007). Removal of the instances in case of missing value may resolve the problem of missing values. However as the number of cases is limited in most real world datasets, the available data for analysis will be reduced. Also by removing such cases, vital information about the data may be lost (Engelbrecht, 2007).

The other method is to replace the missing value with the average value of the remaining available cases in case of continues data, or replacing the missing value with the most occurring value in case of nominal data (Kotsiantis et al., 2006).

It is also possible to estimate the missing value using KNN method (section 1.3.1.1). As discussed previously, KNN can estimate the missing value based on the likelihood of its similar data in the dataset. Research (Kim et al., 2005; Dudoit et al., 2002; Troyanskaya et al., 2001) suggested that KNN may be the most reliable method for estimation of the missing value.

2.3.2 Outliers

In statistical terms, outliers are defined as the data that significantly differs from other data in the dataset (Barnett and Lewis, 1994). Engelbrecht (2007) highlighted the issues of outliers in the data and suggested as the values of outliers are significantly different to the rest of data, it affects the function (or model) that fits to the data and they tend to

pull the resulted function (or model) towards themselves (Figure 2-1). As a result they play an important role in reducing the accuracy of the model.

The problem of outliers can be addressed using different methods. The most common and straightforward approach is to completely remove such data from the data set. It is argued (Engelbrecht, 2007) that removing the outliers from dataset resolves the issues surrounding outliers in the dataset; however it is possible to lose vital information about the data. There are many approaches developed based on removing the outliers. Gedeon et al. (1995) and Slade and Gedeon (1993), developed the approach of removing the outliers during training the model based on the distribution of the data.

Some other researchers (Huber, 1981), suggest that it is possible to build the robust model or function that is not influenced by outliers.

Outliers were included but data were always cross checked between paper and electronic formats and cases with discrepancies discarded. In this research, the removal approach for handling the outliers will be employed. This is because of the nature of the present data for this research, as many of the outliers may have been caused by human error during the entering and recording of the data into paper based files.

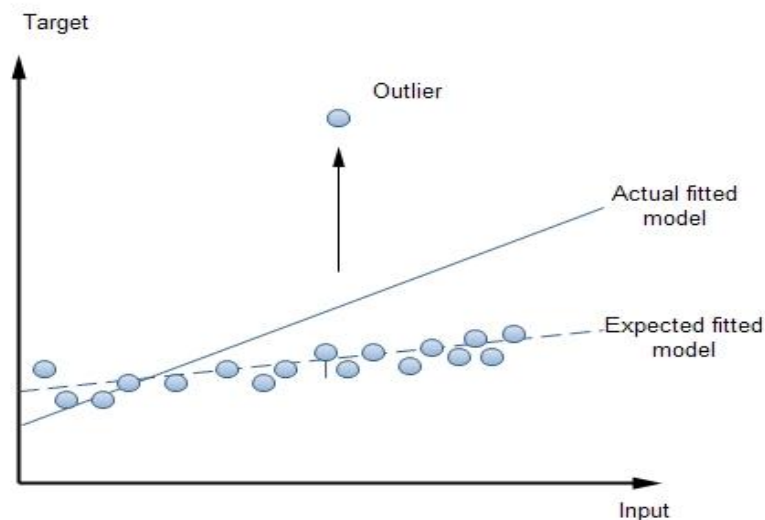


Figure 2-1The issue of outliers in the dataset
(Engelbrecht, 2007)

Outliers tend to distance the desired function (or model) to be closer to them and as a result significantly affect the accuracy of the model.

2.3.3 Coding the input values

Some classifiers and machine learning models such as ANN, can only handle numeric values. Therefore, it is essential to convert nominal values into numeric values before analysis. The solution for the variables that only hold two nominal values such as ‘Sex’ is straightforward, one of the values for example male can be coded to ‘1’ and the other (female) to ‘0’. It is possible to apply the same approach for other variables and map each nominal value to a unique numeric value, however as Engelbrecht (2007) argues, the model may assume that this set of numbers represents a continuous dataset which will cause the model to lose the original characteristic of such data. The possible solution is to use binary labelling for converting the nominal data (Huang and Ng, 2003). The nominal data with n different values can be coded into n different variables where the corresponding variable’s value is set to ‘1’ and the rest of variables are set to ‘0’ (Engelbrecht, 2007). In the other words, the n entry values for a nominal variable will be converted to n different variables that each variable is set to one of the entry values. For each case the new converted variable is set to ‘1’ if and only if it corresponds to the entered value for the nominal variable.

2.3.4. Normalization and scaling

Data normalization and scaling is a common practice which is used for most of the research that involves some sort of data and modelling (Cheng and Li, 2008). It is essential to note that in this thesis, normalization refers to the process of scaling down the values to a specific range, rather than scaling it up. Most of the classifiers and models normalize data before processing. Normalization may ensure higher performance for models and classifiers as this technique preserves the relation between features in the dataset and it simplifies the computing process (Han et al., 2006). It has been argued that the normalization process is an essential part of data pre-processing as it prevents the feature with a large range dominating the other features in the dataset with relatively smaller ranges (Han et al., 2006; Engelbrecht, 2007; Kotsiantis et al., 2006).

There are many methods developed for normalizing the data. One of the most commonly used techniques is to consider the mean and the standard deviation of a feature and normalizing each value of the feature. This method is commonly known as zero-mean normalization. According to Han et al. (2006), consider v is one of the values

of feature X and the standard deviation of X is calculated as σ_x and the mean value of X is \bar{X} . Therefore the v_n normalized value of v can be calculated as:

$$v_n = \frac{v - \bar{X}}{\sigma_x}$$

Equation 2-1 zero-mean normalization

Where V is one of the values of feature X and σ_x is the standard deviation of X and \bar{X} is the mean value of X .

Max-min normalization is another method that is widely used. Assume \max_x and \min_x are the maximum and minimum values of feature X and v one of the values of feature X . v_n the normalized value of v can be calculated as (Han et al., 2006):

$$v_n = \frac{v - \min_x}{\max_x - \min_x}$$

Equation 2-2 max-min normalization

Where v is a value of X , \max_x and \min_x are the maximum and minimum values of feature X and v_n the normalized value of v

There are many other methods developed for data normalization. However the differences between methods are not very noticeable (Han et al., 2006) and due to the simplicity of the zero-mean method, this method is selected for this research.

Although the data normalization is common practice in most research and datasets (Cheng and Li, 2008), we have to keep in mind that in some cases it is essential to consider that some data may show some domination in the data by their nature (such as gene analysis data). Consequently, scaling down the gene expression data may equalise the expression of genes. Therefore, it is essential to consider the data and type of analysis in hand before performing the data normalization.

2.4 Comparison of datasets

In section 2.2 datasets OV_TMA_2000 and OV_TMA_2005 were discussed. Prior to merging the two datasets, they were compared to ensure no significant differences existed between the two. In order to achieve this, a paired sample T test using Wilcoxon, Sign, McNemar and Marginal Homogeneity is performed. These tests are

used to compare groups that are related in some way. The results of these tests indicate the similarity or differences between the values of the features in each dataset (Riffenburgh, 2005).

There is no statistical difference between Age, FIGO, Histology, CA125 and Grade variables in the two datasets. However there was a statistical difference between outcome of the surgery and survival rates.

The comparison of survival and outcome of surgery (Figure 2-2 and Figure 2-3), indicates an increase in the number of complete cytoreduction from 2000-2005 (OV_TMA_2005) compared to 1995-2000 (OV_TMA_2000). Also it shows a decrease in the number of sub-optimal cytoreduction from 1995-2000 compared to 2000-2005. Meanwhile the survival rates increased from 1995-2000 compared to 2000-2005.

In the absence of significant differences between other variables in the datasets, this analysis suggests, a strong correlation between the outcome of surgery and the survival rates of the patients. This may draw a significant conclusion that the attempts to remove as much tumour as possible from the patients significantly increase their survival rate. On the other hand it may be that these days the only patients that considered for surgery are better patients (patients with good physical and mental conditions).

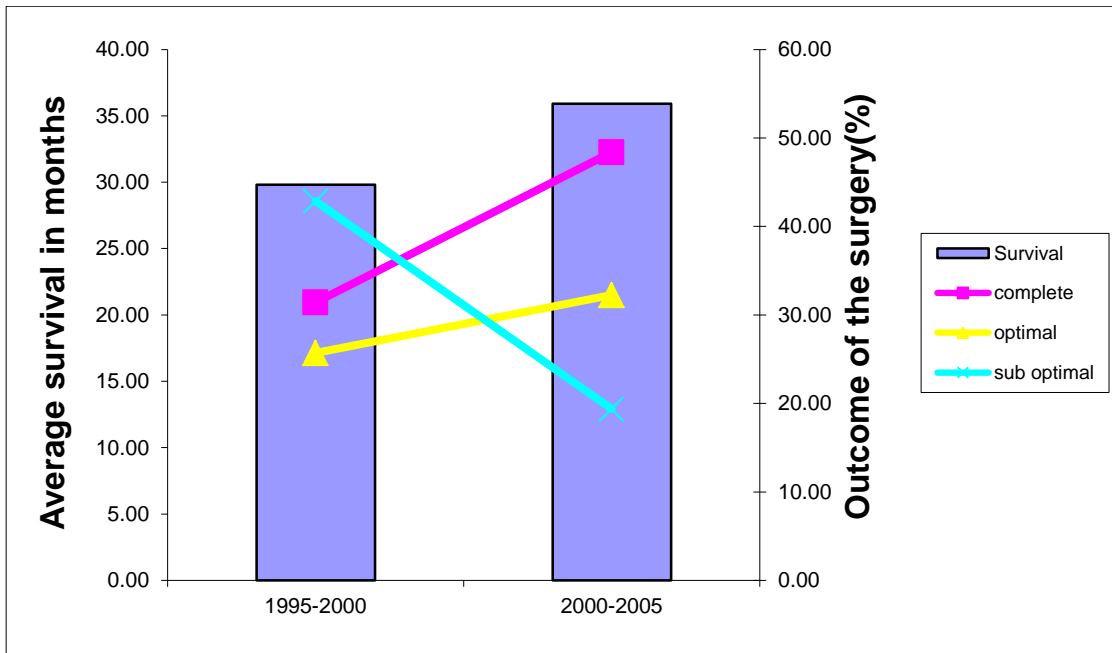


Figure 2-2 Two datasets comparison

The statistical comparison of OV_TMA_2000 and OV_TMA_2005 concluded that there is a statistical significant difference between the outcome of surgery (p value of .006) and survival (p value of .006) in the two datasets.

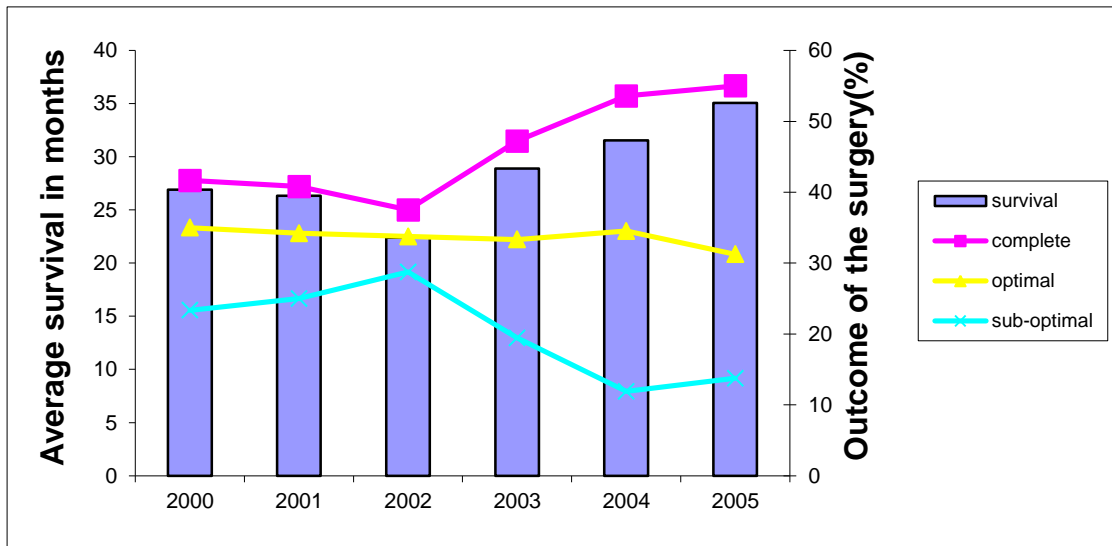


Figure 2-3 Survival vs. Outcome of surgery 2000-2005

Correlation between outcome of surgery and survival rates from 2000-2005.

2.5 Summary of chapter

This chapter of the thesis introduced the datasets that will be used in this research. Each dataset was briefly described and the process of collection was highlighted.

Furthermore, the process of pre-processing of the datasets including the data normalization, handling the missing values and identifying the outliers was described. Following the validation process, the initial comparisons of the two available datasets were presented. This initial comparison highlighted an important conclusion, which is the strong relation between the outcome of the surgery and the survival rate of the patients.

Chapter 3 Experimental methodology

3.1 Introduction

This chapter introduces the methods and metrics used for measuring the performance of the models, which will enable comparison and selection of the best model. There are many estimates and metrics available; however the most popular methods were selected and are briefly described in this chapter. This chapter also introduces the design of algorithms which demonstrates the analysis part of this research. Finally, the tools that were used in this research will be introduced and briefly described.

3.2 Estimates of prediction performance

Machine learning and some of its techniques were discussed previously (section 1.3). Briefly, a typical task in data mining, machine learning or statistics is to train a model using available data, the training data, and then apply the trained model to new data, referred to as evaluation data. As discussed previously (section 1.3) , there are many different algorithms available in this field. Ideally the true risk (performance of all related data) of the model would select the better algorithm. Unfortunately, in real applications only a finite number of data are available, therefore, estimates have to be made (Chapelle et al., 2002). In order to assess the performance of the model, it is essential that the model is not assessed using the data with which it was trained. The model should make appropriate predictions of the evaluation data, (Fujikoshi et al., 2010). The following approaches were proposed to overcome the above issue.

3.2.1 Resubstitution Validation

This approach uses all available data for training the model and evaluates the model using the same data (Braga-Neto et al., 2004). In other words, in this approach data has been used twice, once for training the model and then for evaluation.

The Resubstitution Validation, uses the same data for training and evaluation, and will therefore underestimate the true error rate (Dubitzky et al., 2006). This validation process is widely used in statistics where the number of data instances is relatively larger than the number of parameters (factors) in the dataset, because of its simplicity (Dubitzky et al., 2006). On the other hand, where the number of parameters (factors) in the dataset is large, this estimate tends to be insignificant and misleading (Dubitzky et al., 2006). Simon et al. (2003) argue that to measure the prediction ability of a model the data for training and evaluation should be separate. Figure 3-1 demonstrates the amount of data that is used for training and evaluation of the model.

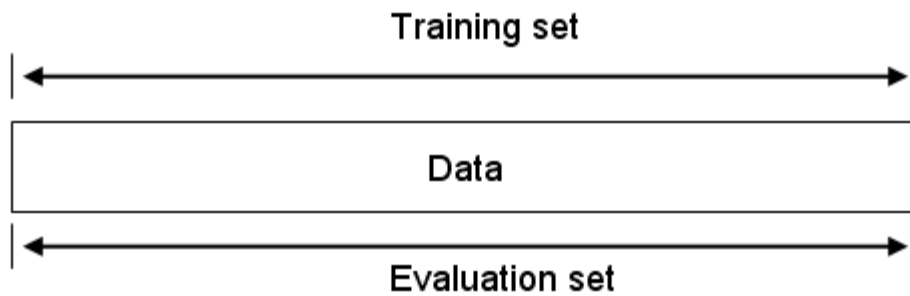


Figure 3-1 Resubstitution Validation

This method uses the entire dataset for training and evaluation.

3.2.2 Hold-out validation

To overcome the problem of the Resubstitution validation method Hold-out validation was introduced. This approach of validation randomly splits the data in to two separate sections, one for training the model and one for evaluation (Lee, 2010). Typically less than one third of the data is selected as the evaluation set (Lee, 2010). In other words the evaluation set part of the dataset is held out and not used during the training process (Figure 3-2). Hold out validation, uses separate data for training and validation, therefore this approach leads us to a more accurate estimate for prediction performance of the model (Lee, 2010).

In this approach, the results of training and validation are highly dependent on the split in the data set (Marzban, 2009). Since this method uses a single training and evaluation over the data set, it may result in an easy or hard section of the data for training and validation and leads to “unfortunate” splits (Marzban, 2009). These problems can be partially addressed by repeating the process multiple times and using the average performance as the final prediction performance. However, as there is no system defined for this process, some of the data may be included in the training set multiple times, while the others are not included at all. It is also possible that some data may always fall into the training set and never have a chance to take part in the evaluation set (Marzban, 2009).

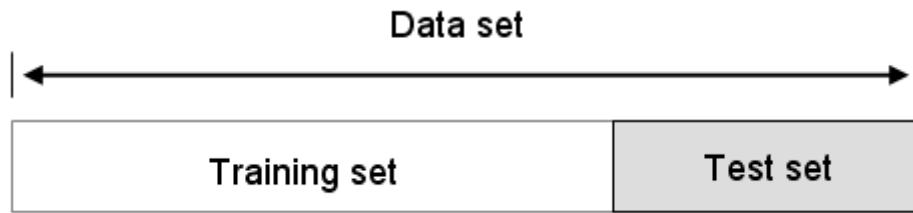


Figure 3-2 Hold-out Validation

Hold-out validation splits the data set into two separate sets for training and validation.

3.2.3 K-Fold Cross-Validation

This method partitions the data into k equally (or nearly equal) sized partition or folds (Figure 3-3). For each of k experiment, the ‘k-1’ folds will be used for training the model and the remaining one for testing (Lee, 2010). In order to make sure that each fold is a good representative of the whole data set, at the start of the process data is “stratified” (Dziuda, 2010). The main disadvantage of this technique is that the training and evaluation have to be repeated k times; therefore it takes k times longer than the two previously described methods. However the training and validation sets are always separate to each other and the whole dataset is covered.

Therefore the true error is estimated as an average error rate on the test sets. Equation 3-1, illustrates the calculation of the true error rate, where k is the number of folds, and E_i is the error rate for each test.

$$E = \frac{1}{k} \sum_{i=1}^k E_i$$

Equation 3-1 True error rate of k-fold cross-validation

Where k is the number of folds, and E_i is the error test for each test.

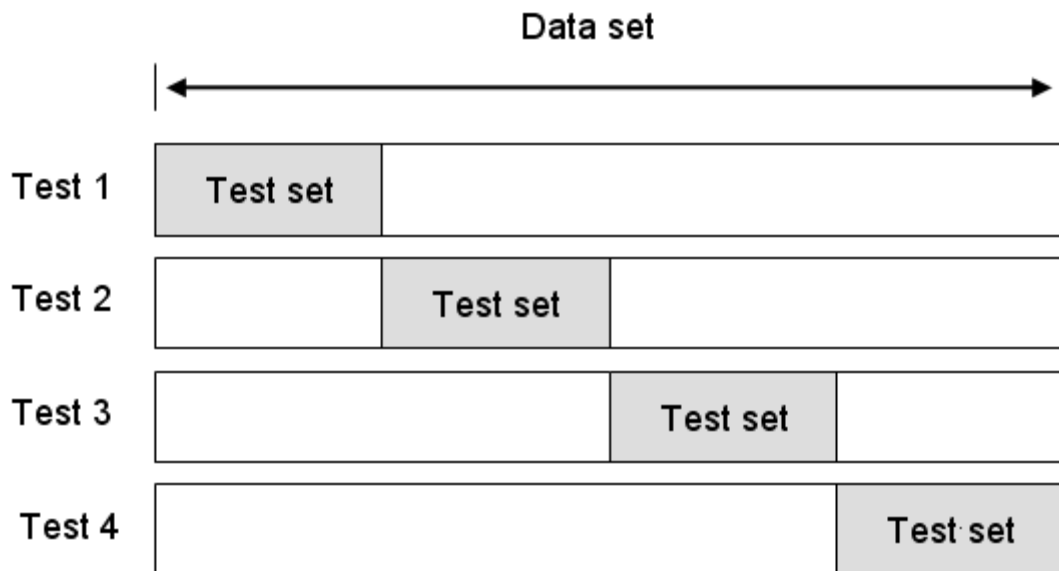


Figure 3-3 4-fold Cross-Validation

4-fold cross-validation splits the dataset into training and evaluation sets. Each time $3(4-1)$ folds will be used for training and the remaining for evaluation. This process will be repeated 4 (k) times.

Most researchers tend to use $k=10$ for cross-validation (Dziuda, 2010). Lee (2010) argues that the number of folds may depend on the size of the available data set. A larger number of folds increase the accuracy of the estimator, on the other hand the variance of the estimator will be large and the computational time will be high due to many experiments. The small number of folds reduces the computational time and the variance of the estimator will be small, however the estimator will be less accurate (Lee, 2010).

3.2.4 Leave-One-Out Cross-Validation (LOOCV)

LOOCV is a special case of k -fold cross-validation, where k is equal to the number of data items in the dataset (Dubitzky et al., 2006). In other words, for a dataset with k number of data, k tests will be performed, where $k-1$ examples will be used for training the model and the remaining 1 example will be used for evaluation (Figure 3-4).

LOOCV is widely used especially where the number of available data is very small

(Dziuda, 2010). The running time in this method is higher than k-fold cross-validation due to increase in number of folds.

The true error rate in this case is calculated similarly to the k-fold cross validation (Equation 3-1).

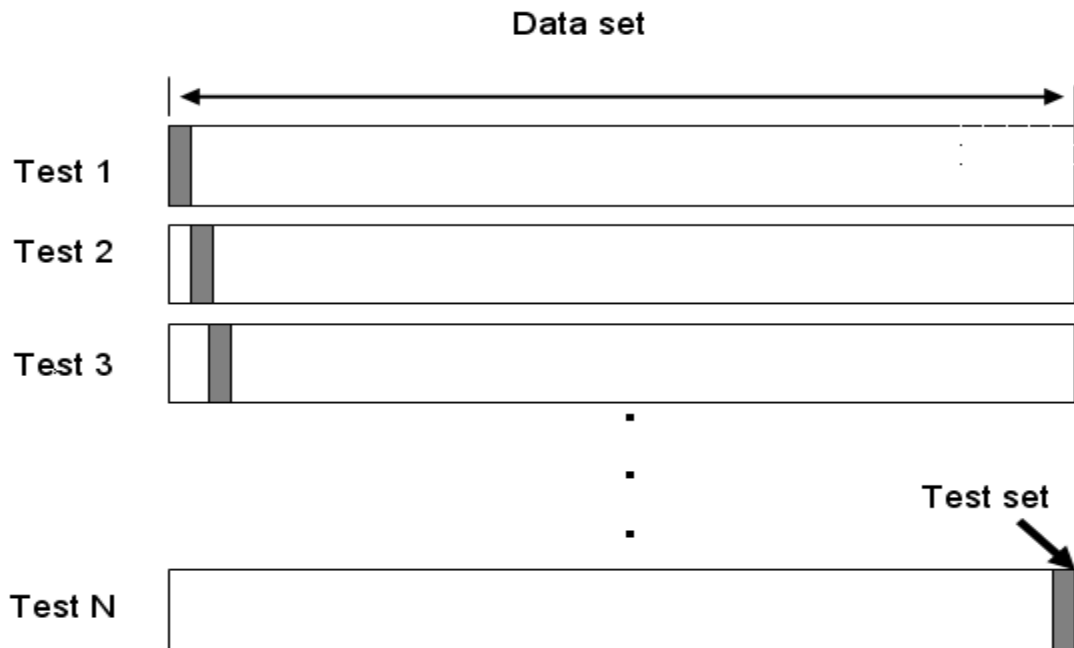


Figure 3-4 Leave-One-Out Cross-Validation

Each time $N-1$ number of data will be used for training and the remaining data for evaluation.

3.2.5 Repeated k-fold Cross-Validation

This method has a similar concept to k-fold cross-validation (Section 3.2.3), and it repeats the process multiple times (Lee, 2010). At the start of each round data is reshuffled.

This method increases the computational time, however Dziuda (2010) argues as the number of estimators' increases it may lead to a higher accuracy.

3.2.6 Evaluation of the estimates

Previously (section 3.2.1-3.2.5) some of the different prediction performance estimators were discussed. There have been many attempts in the past to understand the advantages and drawbacks of each method and conclude which method performances are the best.

Although this conclusion is very dependent to the available data set, it is possible to select one of these methods for estimation of the prediction performance in this research. Kohavi (1995), Salzberg (1997) and Dietterich (1998) compared several methods to estimate accuracy and concluded that 10-fold cross-validation is the best selection method. More recently, Molinaro et al. (2005) compared several methods of estimation on very wide range of datasets and concluded that 10-fold cross-validation performs the best compared to the other methods.

Furthermore many researchers such as Patnaik et al. (2010), Boutros et al. (2009), Hartmann et al. (2009), Dragonieri et al. (2009), Barnett et al. (2010) and Crijns et al. (2009) used one type of cross-validation for their research.

Therefore the 10-fold cross-validation has been selected to estimate the prediction performance in this research.

3.3 Prediction performance measurements

The criteria for evaluating the performance of a model or a classifier are an important part of its design and the research. Using defined criteria allows estimates of the actions of a model or a classifier on unseen data to be calculated and also it can be used to compare the performance of generated models or classifiers. This will help to decide if the model or classifier is good enough for the purpose or to enable the researcher to improve it or replace it with another procedure. Several criteria have been used to evaluate the performance of classifiers or models. This section, introduces some of these criteria.

3.3.1 Possible outcomes of a model

There is a need for measurements to identify which of the used models or classifiers performed better. In order to introduce such measurements, it is essential to define the types of outcomes that a model can produce. Consider a dataset with two classes 1 (“yes”) and 0 (“no”). The following are four possible outcomes of the prediction:

False positive (FP): the outcome incorrectly classified “yes” (or 1 or positive) when the actual class is “no” (or 0 or negative).

False negative (FN): the outcome incorrectly classified as “no”, when the actual class is “yes”.

True positive (TP): the outcome correctly classified as “yes”, when the actual class is “yes”

True negative (TN): the outcome correctly classified as “no”, when the actual value is “no”.

Figure 3-5 summarizes the above predictions:

		Predicted Class	
		Yes	No
Actual Class	Yes	TP	FN
	No	FP	TN

Figure 3-5 The possible outcomes of prediction

The true negative rate is the proportion of class 1 (“yes”) items which are correctly classified as class 1 (“yes”) and it is denoted as tn (Krzanowski and Hand, 2009).

$$tn = 1 - FP$$

Equation 3-2 True Negative Rate

The false negative rate is the proportion of class 0 (“no”) items, which are incorrectly classified as class 1 (“yes”) and is denoted as fn (Krzanowski and Hand, 2009).

$$fn = 1 - TP$$

Equation 3-3 False Negative Rate

3.3.2 Accuracy

It is possible to measure the performance of a model by counting the correctly classified instances and incorrectly classified instances. Furthermore Dobbin et al. (2008) considered the accuracy of a model as number of correctly classified instances in the dataset (Equation 3-4).

$$Accuracy = \frac{\text{number of correctly classified instance}}{\text{number of instances}} = \frac{TP + TN}{TP + TN + FP + FN}$$

Equation 3-4 Accuracy

Where *TP*: true positive, *TN*: true negative, *FP*: false positive, *FN*: false negative

However accuracy on its own will not provide the complete information about the performance of a model. Olson and Delen (2008) suggested that the domination of a category (class) in a dataset affects the value of accuracy and also accuracy returns small information about the performance of the model (for example is 90% accuracy good or bad?). Furthermore Provost et al. (1998) argues that accuracy assumes the equal costs of false positive and false negative; regardless of the fact that in the real world one type of error can be more costly than the other. Therefore Olson and Delen (2008) suggested the following metrics to measure the ‘effectiveness’ of a model or a classifier.

3.3.3 Recall

One of the metrics that is widely used is recall. Recall is the percentage of the correctly identified instances among all instances in the dataset that actually belongs to the given class (Olson and Delen, 2008). Recall is also known as sensitivity or true positive rate of a model. Recall of a given class ‘a’, can be measured as the number of correctly classified instances of that given class over number of instances in that class (a).

$$recall(a) = \frac{\text{number of correctly classified instances of a given class}(s)}{\text{number of instances in class}(a)} = \frac{TP}{TP + FN}$$

Equation 3-5 Recall of given class a

Where ‘a’ is the given class, *TP*: true positive, *FN*: false negative

3.3.4 Precision

This metric is the percentage of the correctly identified instances among instances that the model considered to belong to that class (Olson and Delen, 2008). This can be measured as number of correctly classified instances of given class ‘a’ over number of instances model classified as class ‘a’.

$$precision(a) = \frac{\text{number of correctly classified instances of a given class (a)}}{\text{number of instances model classified as class (a)}} = \frac{TP}{TP + FP}$$

Equation 3-6 Precision of given class a

Where 'a' is the given class, TP: true positive, FP: false positive

When comparing the models, the model with higher value of precision is the better model.

3.3.5 Mean Absolute Error (MAE)

In data mining, mean absolute error, is a metric that shows how the prediction of the model is close to the actual values. Mean absolute error “is the average of the differences between predicted and actual value in all test cases” (Patil et al. , 2010).

Patil et al. (2010) introduces MAE as shown below (Equation 3-7):

$$MAE = \frac{|p_1 - a_1| + \dots + |p_n - a_n|}{n}$$

Equation 3-7 Mean Absolute Error

Where 'p' is the predicted value by the model, 'a' the actual value, n is the total number of instances.

MAE is an error rate; therefore the model with the lower value of MAE is the better model.

3.3.6 Root Mean Squared Error (RMSE)

RMSE is one of the most widely used metrics to calculate the performance of a model (Caipeng et al., 2010). RMSE, like MAE, measures the performance of the model by calculating the difference from the predicted value and the actual value. However as the name suggests the errors are squared, therefore it gives a higher value to large errors. This metric is very useful where large errors are undesirable. According to Patil et al. (2010), RMSE can be measured as below (Equation 3-8):

$$RMSE = \sqrt{\frac{(p_1 - a_1)^2 + \dots + (p_n - a_n)^2}{n}}$$

Equation 3-8 Root Mean Squared Error

Where ‘p’ is the predicted value by the model, ‘a’ the actual value, ‘n’ is the total number of instances.

Similar to MAE, RMSE is an error rate; therefore the model with the lower value of RMSE is the better model.

3.3.7 Receiver Operating Characteristic (ROC)

This method was developed in the 1940s to evaluate the performance of signal detection systems (Zaknich, 2003). More recently the ROC has also been used to analyse the performance of models and classifiers. Previous sections have described the metrics that can be used to measure the performance of the model, and one can use the metric which is suited for the condition. However it is difficult to identify a metric that works in all circumstances and can be used in similar situations in the future. Therefore it is useful to have a method to display and summarise the performance over a wide range of situations. The ROC curve exactly does that (Krzanowski and Hand, 2009).

ROC graph is a plot with true positive rate on the Y axis and false positive rate on the X axis as the classification threshold t varies (Krzanowski and Hand, 2009) (Figure 3-6).

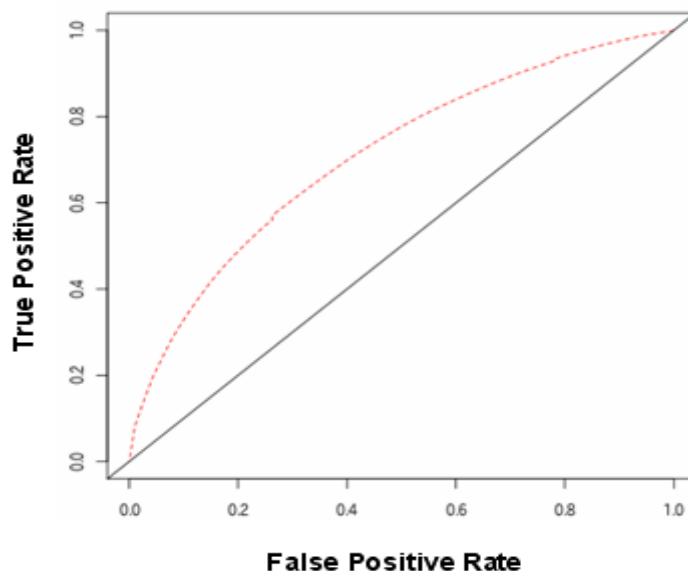


Figure 3-6 Receiver operating characteristic (ROC)

ROC curve plots the true positive rate against the false positive rate and is widely used for the performance analysis of a given model.

In most classifiers or models there is a parameter that can be adjusted to increase the true positive at the cost of a decrease in the false positive or vice versa, this parameter is called the classification threshold or ‘t’. Each value of t provides a (FP, TP) pair that can be plotted on the graph. Justifiably the point (0, 1) represents the perfect model or classifier as it shows that the false positive rate is 0 (or none) and the true positive rate is 1 (or all). The point (0, 0) on the graph represents the model (or classifier) that predicts all instances to be negative, while point (1, 1) corresponds to the model (or classifier) that predicts every instance to be positive.

The areas under the ROC curve which is commonly denoted as AUC is the most widely used summary index (Krzanowski and Hand, 2009).

$$AUC = \int_0^1 y(x)dx$$

Equation 3-9 Area under ROC curve (AUC)

Accordingly the AUC value of 1 represents the perfect test and the value of 0.5 a worthless test. Table 3-1 summarizes the interpretation of AUC values (Krzanowski and Hand, 2009).

AUC	Results
0.90-1.00	Excellent
0.80-0.90	Very Good
0.70-0.80	Good
0.60-0.70	Fair
0.50-0.60	Poor

Table 3-1 AUC results interpretation
(Krzanowski and Hand, 2009)

3.4 Over fitting

Over fitting which is also known as poor generalisation, is a major problem for learning methods and can reduce the accuracy of the algorithm by 10-25% (Mitra and Acharya, 2003). In most cases over fitting occurs where the training data does not adequately represent the entire dataset or there is noise in the dataset (Coppin, 2004). The problem of over fitting may also occur when the number of available data is not enough for a complex model to learn from the data. The number of training data along with the amount of training process is an important factor for a learning model so it can cope with the unseen data. However as Marsland (2009) claims there should be a balance for

the training process, if the model is trained for too long we may over fit the data, as the model should learn about noise in the data and this makes the model more complex. If the model is trained for a short time it may lose its accuracy for predicting the unseen data.

In most situations, an over fitted model will perform well on the training data, but the performance decreases on evaluation data. In other words model T is over fitted where there is another model M that gives a higher error on the training data, while giving less error on evaluation data (Mitra and Acharya, 2003).

However over fitting can be used as a tool to discover when the training process has to end. The most obvious option is to set the N number of repetitions and stop the training when the number N is reached. However in this case the model may not learn sufficiently from the data or conversely the model has already become over fitted by then. The other solution is to continue the training until the model reaches a threshold error rate. This solution also has its draw backs as the model may never reach the defined error rate or over fits.

The possible solution is to somehow combine the two above approaches together.

Marsland (2009) suggested that the error rate during the training phase along with the validation error rate during the validation phase may provide us with such a tool.

If we plot the error rates during the training phase (which will be reduced over the training period until it reaches a local minimum) along with the validation performance it gives us an indication on when to stop training. Validation error decreases, until a point when it escalates again (Figure 3-7). This point is the over fitting point, as the model starts to include noise in the model.

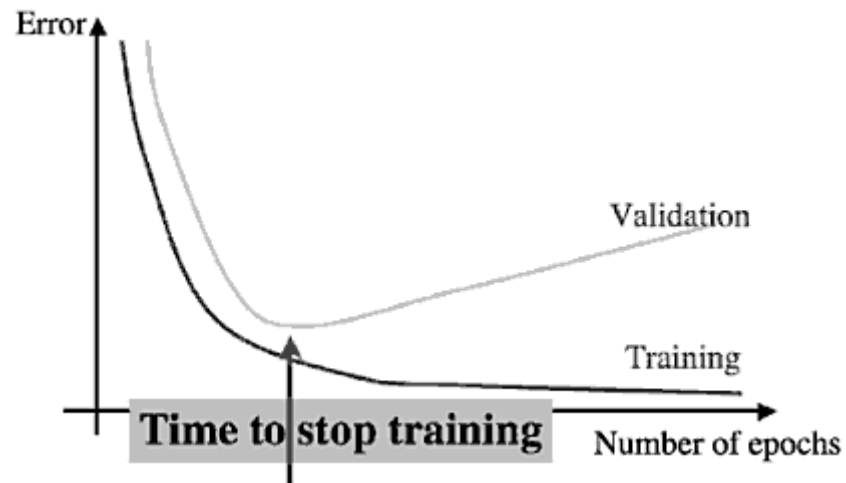


Figure 3-7 Stop training time

Marsland (2009)

The plot of error rate during the training phase and the error rate during the validation phase provides us with a stopping point for training.

3.5 Process definition

A defined process definition is needed for assessing the discussed machine learning (section 1.3.1) and datasets (section 2.2) to discover the highest performance method for constructing the final model. The following diagram (Figure 3-8) explains the process of assessing the models which is generated during this research:

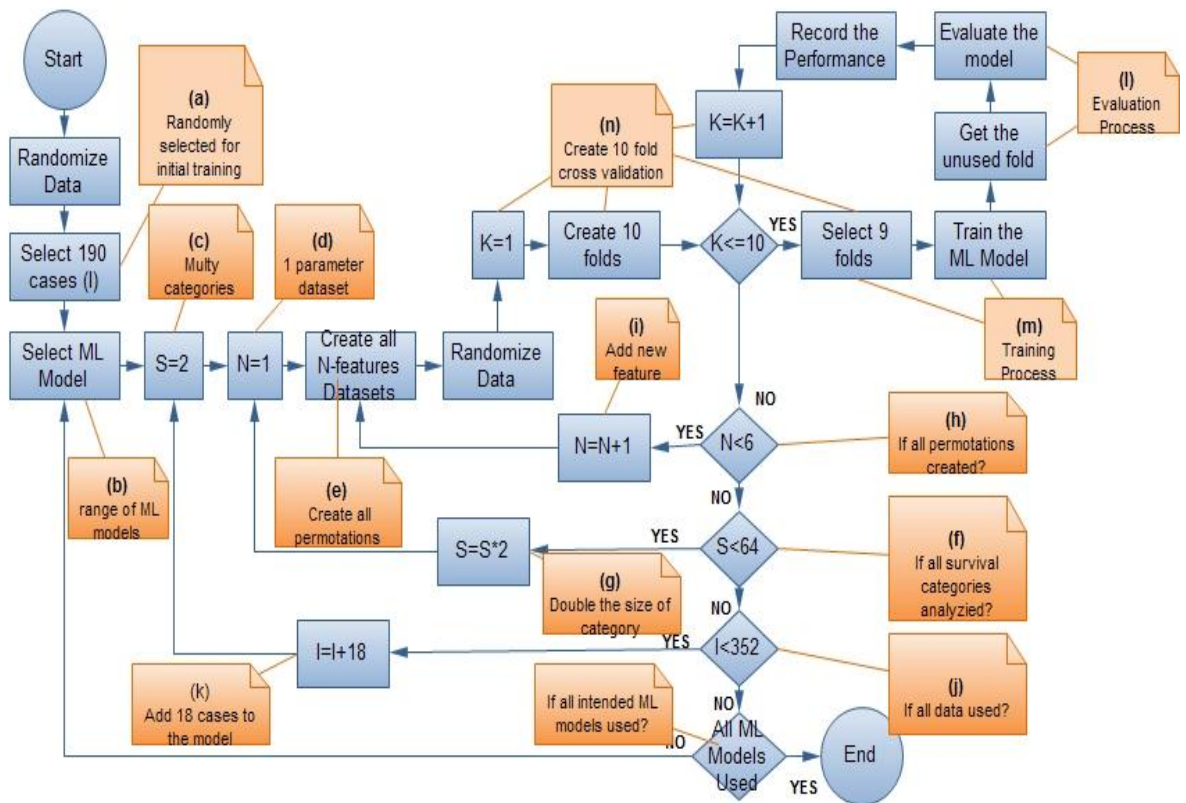


Figure 3-8 Process definition

- At the start of the process the dataset is randomized. The initial training and evaluation of the model starts with 190 cases randomly selected from the dataset (a). Each time 18 new cases are added to the dataset (k) until all 352 available cases are added to the dataset (j). This process will also enables investigation of the model's performance by increasing the number of data and will help us discover when to stop training the model (as discussed in section 3.4).
- To discover the best suited machine learning method for this research, different methods will be assessed (b).
- In order to predict survival models improved to categorize the survival rate for different categories (c). The highest survival rate in the mentioned dataset was 116 months. At the start the system calculates the median point for the survival in the particular dataset and predicts the survival rate as less/more than median (variable C). Then each category will be divided into half and the prediction performance calculated and recorded (g) and the process will be continued until the survival rate is divided in 64 categories (f).
- In order to find the best feature set which maximises the performance of the classifiers, all possible combinations and permutations of the features were investigated (d). The process starts by creating the datasets that only contain one feature (variable N) and it continues to create all permutations of the features (e). Each time a new feature will be added to the dataset (i), until all 6 available features were assessed (h). It is possible to predict this feature set using algorithms that was discussed in chapter 1 (section 1.3.2), however as the size of the dataset is relatively small and the computing power is available, this assessment may provide us with a bench mark for comparing the performance of feature selection techniques.
- Each interval of the process contains a training (m) and evaluation (l) process. The 10-fold cross validation (as discussed in section 3.2.3) is used to estimate the performance prediction (n).

3.6 Summary of the chapter

This chapter introduced the estimates of the prediction performance of a model. These estimates are an essential part of the research as each one introduces a view point for building a frame work for measuring the performance of the model. Based on the evidence discussed, it was decided to use the k-fold cross validation.

This chapter also introduced the prediction performance measurements. These are the metrics that made possible a comparison of the models and selection of the best possible one. All the metrics that were briefly discussed in this chapter will be used to measure the performance of the models. Although ROC analysis is one of the most popular and powerful methods, all of the other metrics also provide us with essential information.

The design and a brief description of the analysis process for this research are included in this chapter.

Chapter 4 Prognostic Data

4.1 Introduction

This chapter investigates the possibility of predicting the survival rate of an ovarian cancer patient. The most important features in the dataset will be introduced and the optimum feature set size will be discovered. The survival rates will be divided into many categories and the prediction performance for each survival category will be analysed. The effects of increasing the number of data items on prediction performance will be discussed and analysed. Furthermore the best possible model for this research will be discovered. The prediction results of the best model will be compared to the prediction results of a conventional statistical model. Finally, previously developed methods will be briefly described and the prediction performance of such models will be compared to the best developed model in this research.

4.2 Finding the best feature set

The process of finding the best feature set has been discussed previously (section 3.5). Briefly, this process enabled the model to investigate all the possible combinations and permutations of the features. This analysis highlights the most important feature and feature sets amongst all other feature and feature sets. As standard practice, to perform this analysis, feature selection techniques are used. However, the relatively small dataset and the available computer power for this research provided us with the possibility to establish a bench mark for comparison of the feature selection techniques ability. A total number of 7,668 analyses were performed. This investigation may answer two important questions. Firstly, what is the optimum size (less than six) feature set that can be used? Secondly, what are the most important features in the dataset?

4.2.1 The optimum feature set size

As discussed previously, all possible permutations of the features were analysed.

Figure 4-1 illustrates the percentage that each feature set achieved the highest accuracy (based on all the performed analysis). The combination of two features in most cases (33.93% of all analyses) achieved the highest accuracy. The combination of three features is in second place (29.46% of all analyses), and in third place is the one feature sets (20.09% of all analyses).

Figure 4-2 illustrates the percentage of total number of analysis that each feature set achieved the lowest value of MAE. The combination of three features in most cases (31.74% of all analyses) achieved the lowest value of MAE. The combination of four

features in second place (25.22% of all analyses) and in third place the combination of two features (22.94% of all analyses) achieved the lowest value of MAE.

Figure 4-3 illustrates the percentage of total number of analysis that each feature set achieved the lowest value of RMSE. The combination of three features achieved the lowest value of RMSE (37.94% of all analyses). The combination of two features in second place (36.05% of all analyses) and in third place one feature sets (18.32% of all analyses) achieved the lowest value of RMSE.

The combination of two features achieved the highest number of accuracy among all analyses (33.93% of cases) (Figure 4-1). However the results of MAE (22.94% of total analyses) and RMSE (36.05% of all analyses) put the combination of two features in second place (Figure 4-2 and Figure 4-3), while the combination of three features was in second place in terms of accuracy (29.46% of all analyses) and the best results of MAE (31.74% of all analyses) and RMSE (37.94% of all analyses).

Considering the results of accuracy (Figure 4-1) and the results of MAE (Figure 4-2) and RMSE (Figure 4-3), it can be concluded that the optimum feature set size is the combination of three features, in this dataset.

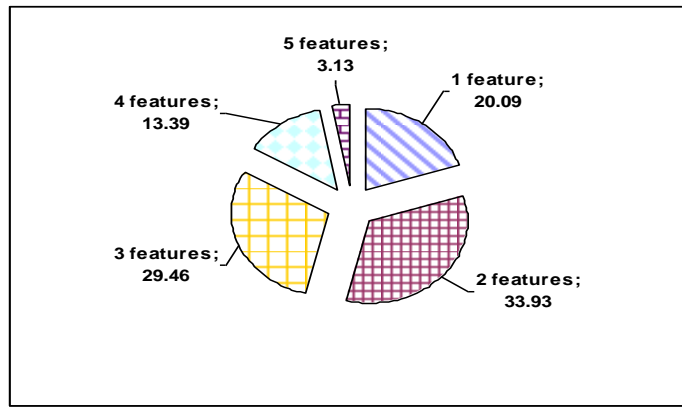


Figure 4-1 the optimum feature set size (Accuracy)

The percentage of total number of analysis that each feature set achieved the highest accuracy.

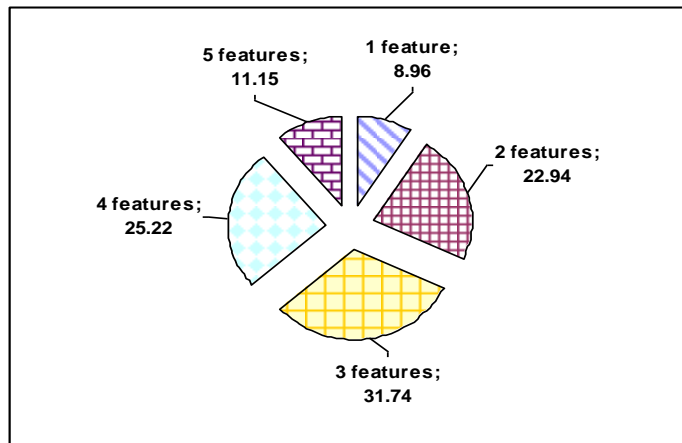


Figure 4-2 the optimum feature set size (MAE)

The percentage of total number of analysis that each feature set achieved the lowest value of MAE.

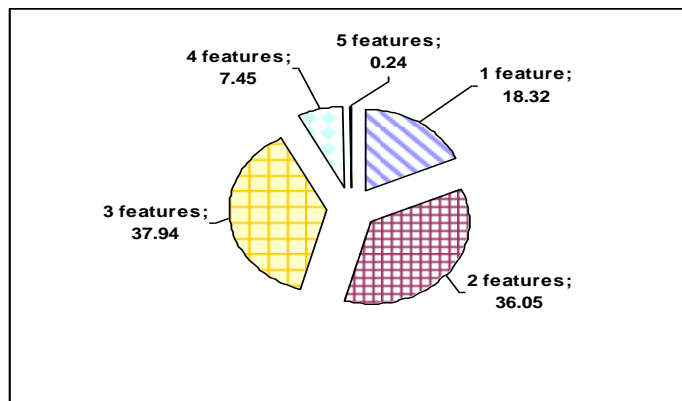


Figure 4-3 the optimum feature set size (RMSE)

The percentage of total number of analysis that each feature set achieved the lowest value of RMSE.

4.2.2 The most important features in the dataset

In order to discover the most important features in the dataset, all features set sizes were investigated. The combination of features that in most number of analyses achieved the highest value of accuracy and lowest values of MAE and RMSE were recorded. In some analysis more than one feature set achieved the same highest performance results (or lower values of MAE and RMSE). Therefore, it was essential to present the values in terms of numbers to address this situation.

Based on the number of times a feature set achieved the highest values of prediction performance and lowest values of MAE and RMSE the best feature set is selected.

The results of these investigations are as follows:

4.2.2.1 One feature sets

Figure 4-4 a: In most number of cases Age and Histology achieved the highest performance (129 times). In third place the outcome of the surgery (123 times) and in fourth place FIGO (107 times).

Figure 4-4 b: Grade achieved the lowest MAE in most cases (134 times). In second place outcome of the surgery (126 times) and in third place CA125 (101 times).

Figure 4-4 c: Grade in most number of cases achieved the lowest RMSE (130 times). In second place outcome of the surgery (124 times) and in third place CA125, Age and FIGO (96 times).

Age in most number of analyses ranked number one in terms of accuracy. However Age achieved the lowest number of times that MAE is the lowest and fourth place in terms of lowest values of RMSE. Meanwhile Outcome of the surgery was second place in terms of accuracy, MAE and RMSE.

Based on the results it can be concluded that the Outcome of the surgery is the most important feature if a single feature dataset was selected.

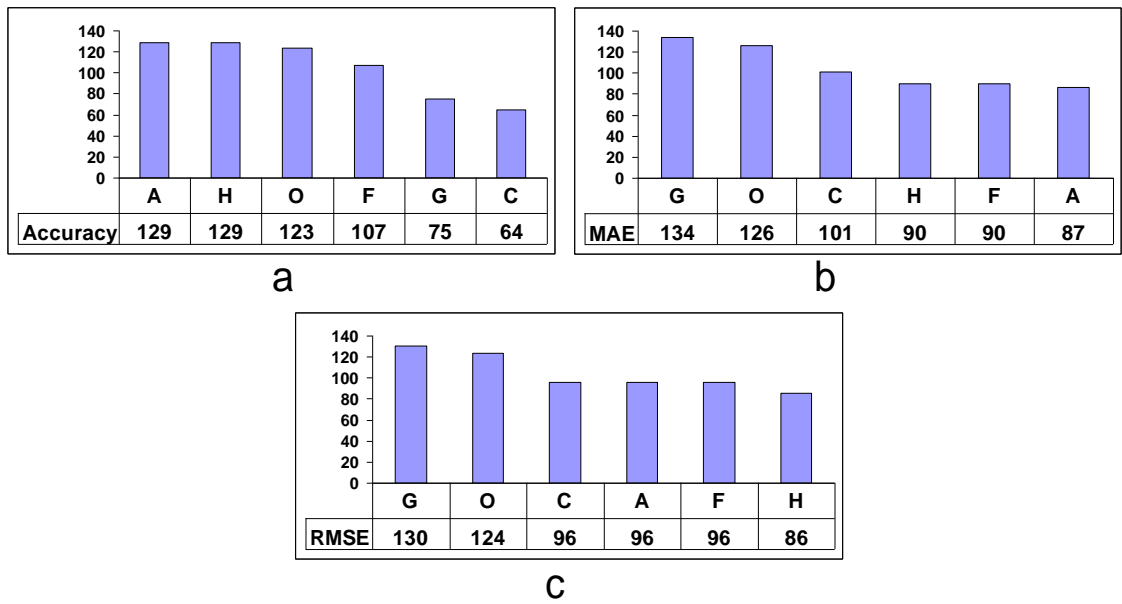


Figure 4-4 one feature sets performance results

a: Accuracy, b: MAE, c: RMSE

A: Age, F: FIGO, G: Grade, H: Histology, O: Outcome of surgery and C: CA125

Y Axes: The number of times each set achieved the highest values of accuracy or lowest values of MAE or RMSE.

4.4.2.2 Two features sets

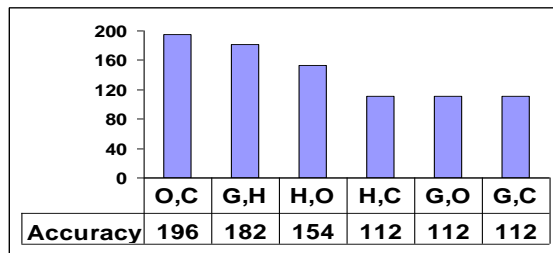
Figure 4-5 a: The combination of Outcome of the surgery and CA125 achieved the highest accuracy in most number of cases (196 times). The combination of Grade and Histology in second place (182 times) and in third place the combination of Histology and Outcome of the surgery (154 times).

Figure 4-5 b: The combination of Grade and Outcome of the surgery in most number of cases (130 times) achieved the lowest values of MAE. In second place the combination of Outcome of the surgery and CA125 (126 times) and in third place the combination of Age and Grade (122 times).

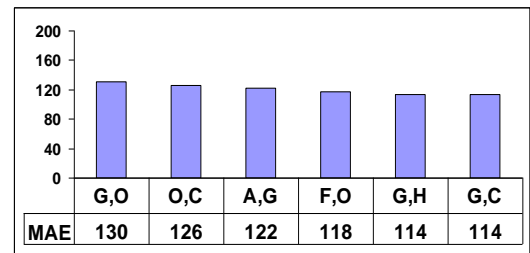
Figure 4-5 c: The combination of Grade and CA125 as well as Grade and Outcome of the surgery in most number of cases (131 times) achieved the lowest values of MAE. In second place the combination of Outcome of the surgery and CA125 (127 times) and in third place the combination of Age and Grade (118 times).

The combination of Outcome of the surgery and CA125 performed the best in terms of accuracy. This combination was also in second place in terms of the lowest values of MAE and RMSE. The combination of Grade and Histology performed reasonable in terms of accuracy; however this combination did not achieve good scores in terms of MAE and RMSE.

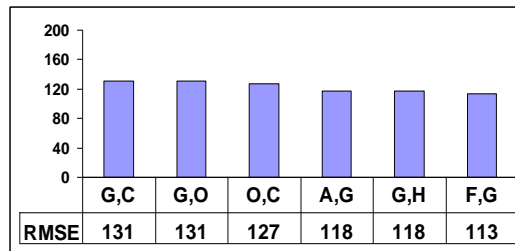
Based on the results it can be concluded that the combination of Outcome of the surgery and CA125 is the most important feature set, if a two features dataset was selected.



a



b



c

Figure 4-5 two features sets performance results

a: Accuracy, b: MAE, c: RMSE

A: Age, F: FIGO, G: Grade, H: Histology, O: Outcome of surgery and C: CA125

Y Axes: The number of times each set achieved the highest values of accuracy or lowest values of MAE or RMSE.

4.4.2.3 Three features sets

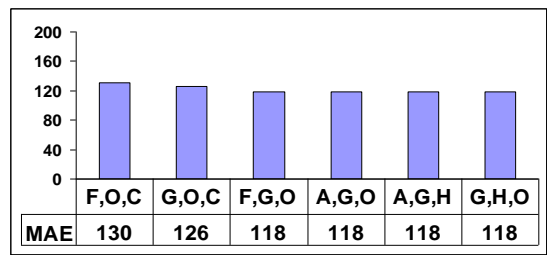
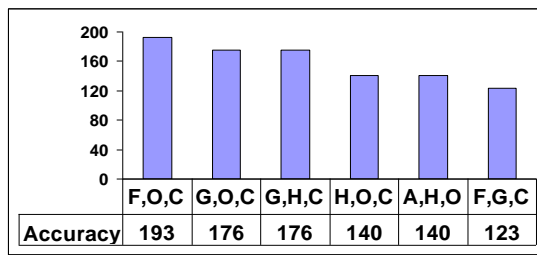
Figure 4-6 a: The combination of FIGO, Outcome of the surgery and CA125 in most number of cases (193 times) achieved the highest values of accuracy. In second place the combination of Grade, Outcome of the surgery and CA125 (176 times) as well as the combination of Grade, Histology and CA125 (176 times). In third place the combination of Histology, Outcome of the surgery and CA125 (140 times).

Figure 4-6 b: The combination of FIGO, Outcome of the surgery and CA125 in most number of cases (130 times) achieved the lowest values of MAE. In second place the combination of Grade, Outcome of the surgery and CA125 (126 times).

Figure 4-6 c: The combination of Grade, Outcome of the surgery and CA125 in most number of cases (140 times) achieved the lowest values of RMSE. In second place the combination of FIGO, Outcome of the surgery and CA125 (123 times) and in third place the combination of Age, Grade and CA125 (119 times).

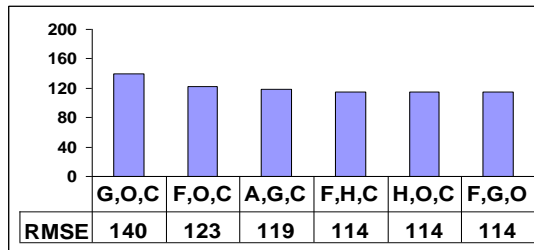
The combination of FIGO, Outcome of the surgery and CA125 achieved the best results in terms of the accuracy and MAE values. This combination scored second in terms of RMSE values. The combination of Grade, Outcome of the surgery and CA125 was closely behind and scored second place overall.

Based on the results it can be concluded that the combination of FIGO, Outcome of the surgery and CA125 is the most important feature set, if a three features dataset was selected.



a

b



c

Figure 4-6 three features sets performance results

a: Accuracy, b: MAE, c: RMSE

A: Age, F: FIGO, G: Grade, H: Histology, O: Outcome of surgery and C: CA125

Y Axes: The number of times each set achieved the highest values of accuracy or lowest values of MAE or RMSE.

4.4.2.4 Four features sets

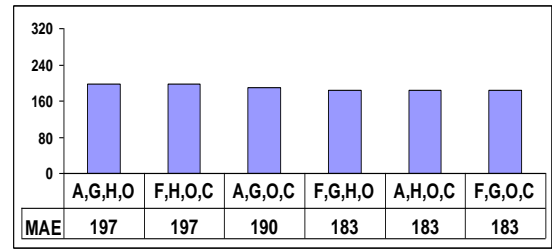
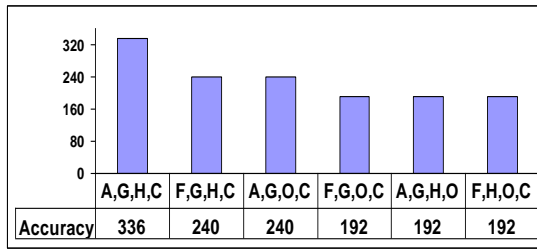
Figure 4-7 a: The combination of Age, Grade, Histology and CA125 in most number of cases (336 times) achieved the highest results of accuracy. In second place the combination of FIGO, Grade, Histology and CA125 (240 times) and the combination of Age, Grade, Outcome of the surgery and CA125 (240 times).

Figure 4-7 b: The combination of Age, Grade, Histology and Outcome of the surgery in most number of cases (197 times) and the combination of FIGO, Histology, Outcome of the surgery and CA125 achieved the lowest values of MAE. In second place the combination of Age, Grade, Outcome of the surgery and CA125 (190 times).

Figure 4-7 c: The combination of FIGO, Grade, Outcome of the surgery and CA125 in most number of cases (306 times) achieved the lowest values of RMSE. In second place the combination of Age, Grade, Outcome of the surgery and CA125 (293 times).

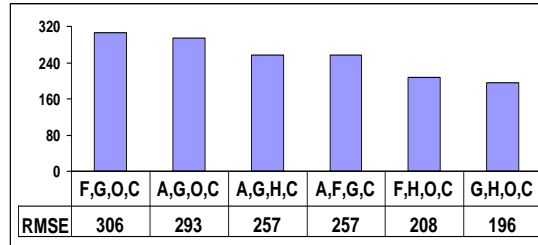
The combination of Age, Grade, Outcome of the surgery and CA125 is in second place in terms of the values of accuracy, MAE and RMSE. All the other combinations failed to achieve the high values of accuracy and low values of MAE and RMSE.

Based on the results it can be concluded that the combination of Age, Grade, Outcome of the surgery and CA125 is the most important feature set, if a four features dataset was selected.



a

b



c

Figure 4-7 four features sets performance results

a: Accuracy, b: MAE, c: RMSE

A: Age, F: FIGO, G: Grade, H: Histology, O: Outcome of surgery and C: CA125

Y Axes: The number of times each set achieved the highest values of accuracy or lowest values of MAE or RMSE.

4.4.2.5 Five features sets

Figure 4-8 a: The combination of Age, FIGO, Grade, Histology and CA125 in most number of cases (202 times) achieved the highest accuracy. In second place the combination of Age, Grade, Histology, Outcome of the surgery and CA125 (187 times) and in third place the combination of FIGO, Grade, Histology, Outcome of the surgery and CA125 (101 times).

Figure 4-8 b: The combination of Age, Grade, Histology, Outcome of the surgery and CA125 in most cases (118 times) achieved the lowest values of MAE. In second place the combination of Age, FIGO, Grade, Histology and outcome of the surgery (114 times) and in third place the combination of Age, FIGO, Histology, Outcome of the surgery and CA125 (109 times).

Figure 4-8 c: The combination of Age, FIGO, Histology, Outcome of the surgery and CA125 in most number of cases (123 times) achieved the lowest values of RMSE. In second place the combination of Age, FIGI, Grade, Outcome of the surgery and CA125 (119 times) and the combination of Age, Grade, Histology, Outcome of the surgery and CA125 (119 times).

The combination of Age, Grade, Histology, Outcome of the surgery and CA125 was in second place in terms of accuracy, number one in terms of MAE and number two in terms of RMSE. The combination of Age, FIGO, Grade, Histology and CA125 achieved the highest score of accuracy and third place in terms of MAE and RMSE values.

Based on the results it can be concluded that the combination of Age, Grade, Histology, Outcome of the surgery and CA125 is the most important feature set, if a five features dataset was selected.

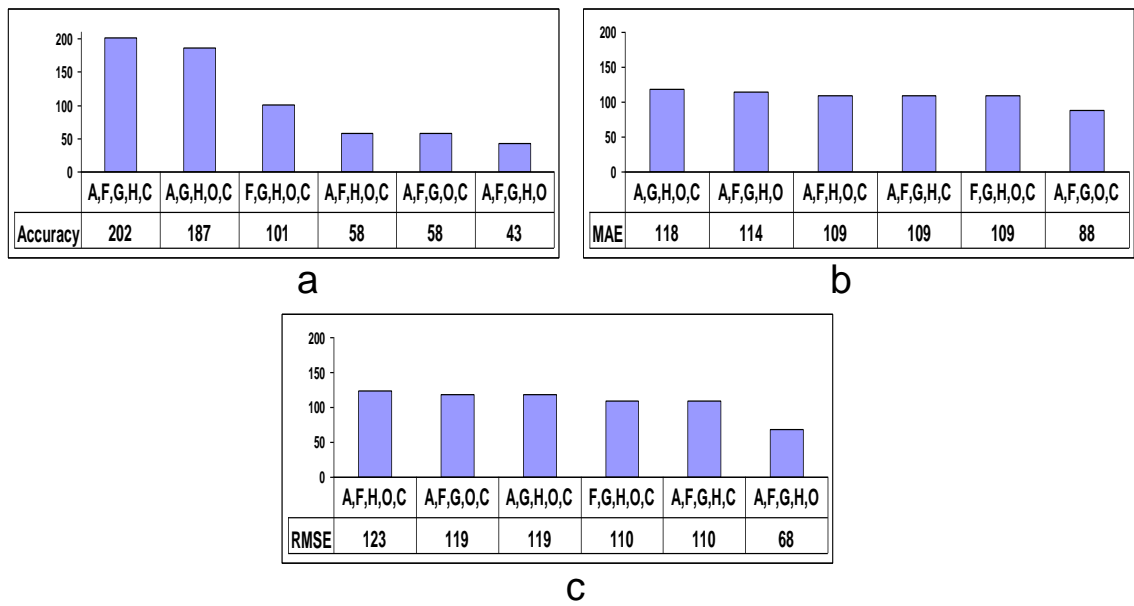


Figure 4-8 five features sets performance results

a: Accuracy, b: MAE, c: RMSE

A: Age, F: FIGO, G: Grade, H: Histology, O: Outcome of surgery and C: CA125

Y Axes: The number of times each set achieved the highest values of accuracy or lowest values of MAE or RMSE.

4.4.3 Discovering best features using feature selection techniques

The feature selection techniques were discussed previously (section 1.3.2). Briefly, the role of these techniques is to identify the most important features (factors) in the dataset.

The most important factors were discovered and discussed previously. These results can be compared against the results of feature selection techniques. This comparison can identify the best feature selection technique among the discussed techniques.

Table 4-1 summarizes the results of using feature selection techniques for identification of the most important features. TPP identified Age, FIGO and Outcome of the surgery as the most important features. IG discovered Grade, FIGO and Age as the most important features. Grade, Histology and CA125 selected by PCA and GA highlighted the importance of FIGO, Age and Outcome of the surgery.

Earlier in this section the most important features were identified as FIGO, Outcome of the surgery and CA125. Based on the results of using feature selection techniques (Table 4-1), none of the techniques completely discovered the best possible feature set. However TPP and GA generated the same results.

All of the techniques failed to discover the optimum feature set; however GA and TPP performed better compared to two other techniques.

Technique	Selected feature
TPP	Age, FIGO, Outcome of surgery
IG	Grade, FIGO, Age
PCA	Grade, Histology, CA125
GA	FIGO, Age, Outcome of surgery

Table 4-1 Important features using feature selection techniques

TPP: Targeted projection pursuit, IG: Information Gain, PCA: Principal Components Analysis, GA: Genetic Search

4.3 Prediction of survival for many categories

Later on (section 4.4), the potential problems and benefits of categorization of continuous data are discussed. However in real world applications it may be beneficial to compromise accuracy to achieve better response. In this research, it is more convenient for a clinician to predict the survival of patients in categories instead of the actual survival value. This will enable the relevant person to predict the domain (or category) of the survival.

In all analyses, the tolerable level is the accuracy level as close as possible to 100% and also the AUC more than 0.70. The error rate also has to be no more than 0.25.

4.3.1 Analysis of 64 survival categories

In order to predict survival of each case, models were generated to categorize the survival data in categories. The highest survival rate in months in the mentioned dataset was 116 months. The system at the start calculates the median point for the survival numbers in the particular dataset and predicts the survival rate as less/more than median. Then each category will be divided in half and the prediction performance calculated and the process will be continued until the survival rate is divided in 64 categories. Based on the survival categories model the prediction categories are as following:

2 categories: 1- 58 months and 58 to 116 months.

4 categories: 1-29 months, 29-58 months, 58-87 months and 87-116 months.

8 categories: 1-14.5 months, 14.5-29 months ... 87-101.5 months and 101.5- 116 months.

16 categories: 1-7 months, 7-14.5 months ... 101.5-109 months and 109-116 months.

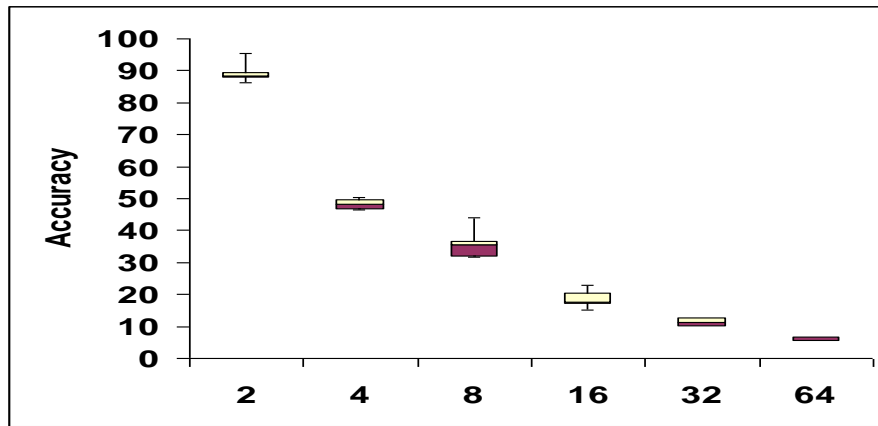
32 categories: 1-3.5 months, 3.5-7 months ... 109-112.5 months and 112.5-116 months.

64 categories: 1-2 months, 2-4 months ... 112-114 months and 114-116 months.

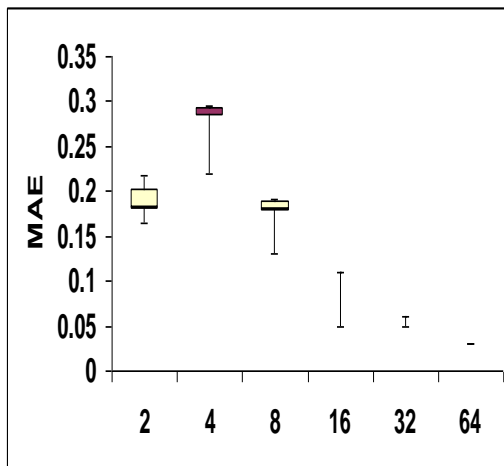
Figure 4-9 demonstrates the prediction performance results for 64 survival categories. Figure 4-9 a, illustrates the prediction accuracy for 64 survival categories. According to these results the prediction accuracy for two survival categories (less or more than median) was around 90%. As expected, as the number of categories increased, the accuracy decreased. These results also demonstrated that the prediction accuracy for 64 survival categories was around 10%. Figure 4-9 b, illustrates the MAE results. The MAE value for two survival categories was around 0.18 and for 64 survival categories

was around 0.03. The MAE value decreased as the number of categories increased in most cases (the only exception was for four categories). Figure 4-9 c, illustrates the results of RMSE. The RMSE value for two survival categories was around 0.31 and for 64 survival categories was around 0.12. The RMSE value decreased as the number of categories increased in most cases (the only exception was for four categories).

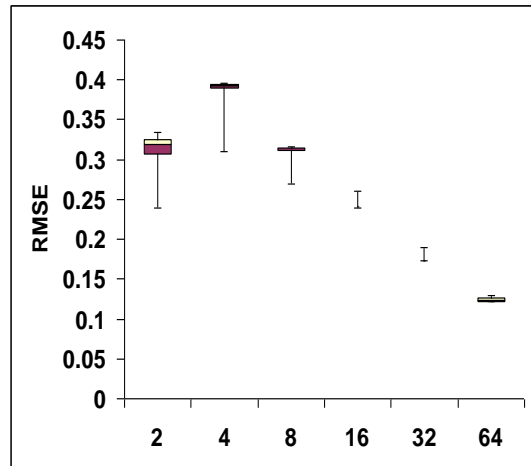
Table 4-2 summarizes the prediction performance accuracy for 64 survival categories. These results were collected and recorded for all the analyses that performed on the dataset. According to these results, model could predict the survival categories of less or more than median with 93.64% of accuracy. As expected the accuracy decreased as the number of categories increased. As a result the prediction accuracy for 64 survival categories recorded as 9.21%.



a



b



c

Figure 4-9 Prediction performance results for 64 categories of survival
(a): Accuracy,(b): MAE and (c): RMSE

X axes: number of survival categories

Number of categories	2	4	8	16	32	64
Minimum	87.09	45.99	30.20	16.83	9.84	3.65
Median	88.34	48.39	35.39	17.77	11.30	6.54
Maximum	93.64	52.37	42.55	21.21	19.95	9.21

Table 4-2 Prediction performance results for 64 survival categories (Accuracy)

Figure 4-10 illustrates the ROC curves for the best model when the number of categories was two. Figure 4-10 a, is the ROC curve for survival prediction less than the median value. The AUC for this curve was 0.72. Figure 4-10 b, is the ROC curve for survival prediction more than the median value. The AUC for this curve was 0.72.

Figure 4-11 illustrates the ROC curves for the best model when the number of categories was four. Figure 4-11 a, is the ROC curve for the survival prediction when the survival was between 0 and 29 months. The AUC for this curve was 0.63. Figure 4-11 b, is the ROC curve for the survival prediction when the survival was between 29 and 58 months. The AUC for this curve was 0.60. Figure 4-11 c, is the ROC curve for survival prediction when the survival was between 59 and 87 months. The AUC for this curve was 0.60. Figure 4-11 d, is the ROC curve for survival prediction when the survival was between 87 and 116 months. The AUC for this curve was 0.61.

The model could predict the survival rates for many survival categories. The accuracy of prediction for two survival categories (less or more than median) was around 90%. The final model predicted the survival categories by 93.64% accuracy where the prediction was based on less or more than the median value. These predictions recorded relatively small MAE and RMSE values. As expected the accuracy of the model decreased as the number of categories increased. As the result the accuracy of the prediction recorded as 9.21% for 64 survival categories.

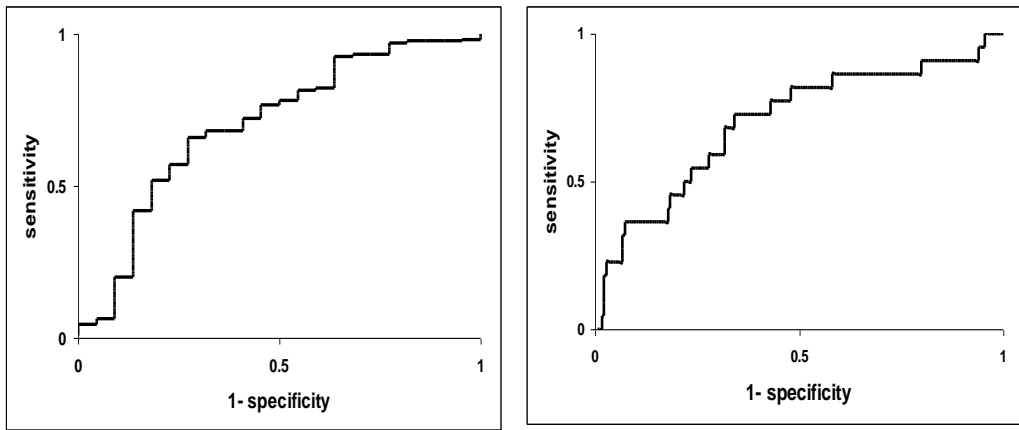


Figure 4-10 ROC curves: 2 survival categories

(a): survival is less than the median value. (b): survival is more than the median value

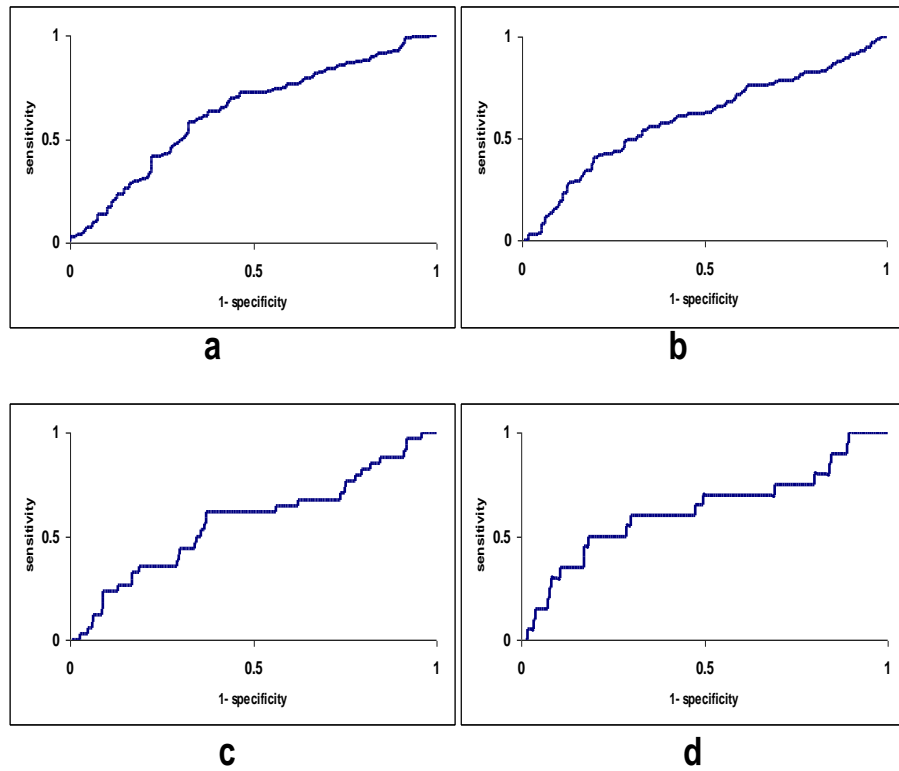


Figure 4-11 ROC curves: 4 survival categories

*(a): survival between 0 and 29 months; (b): survival between 29 and 58 months;
(c): survival between 58 and 87 months ;(d): survival between 87 and 116 months.*

4.3.2 Analysis of Overall Survival (OS)

Previous section (section 4.3.1) analysed and discussed the prediction of survival for 64 categories. As mentioned the highest survival in months in dataset was 116 months. The first analysis divided dataset into two categories: survival less or more than 58 (median value). However, as standard practice most researchers are interested about Overall Survival (OS) of five years (or 60 months). Therefore to perform the survival prediction in a standard manner and also to make comparison with other methods easier, the model was also adjusted to predict the five year survival.

Table 4-3 summarizes the results of OS prediction performance for the best model. The accuracy of the prediction is 92.7 %. The error rate values are relatively small: MAE: 0.16 and RMSE: 0.30. The recall values for both categories (less/more than 5 years) are 0.25 and 0.30 respectively. The precision values for both categories are high, 0.89 for less than 5 years and 0.82 for more than five years. The area under the ROC curve (AUC) indicated that the model performed very good. The AUC values for less/more than 5 years were 0.74.

Figure 4-12 illustrates the ROC curves for OS prediction. Figure 4-12 a, is the ROC curve for less than five years category. The AUC for this curve was 0.74. Figure 4-12 b, is the ROC curve for more than five years category. The AUC for this curve was 0.74.

The produced model can predict the OS by 92.07% accuracy. The error rates of this model were small and the AUC results were 0.74 for both categories.

Measurement		Result	
Accuracy		92.07%	
Mean absolute error		0.16	
Root mean squared error		0.30	

Outcome	Recall	Precision	AUC
Less than 5 years	0.25	0.89	0.74
More than 5 years	0.30	0.82	0.74

Table 4-3 OS prediction performance

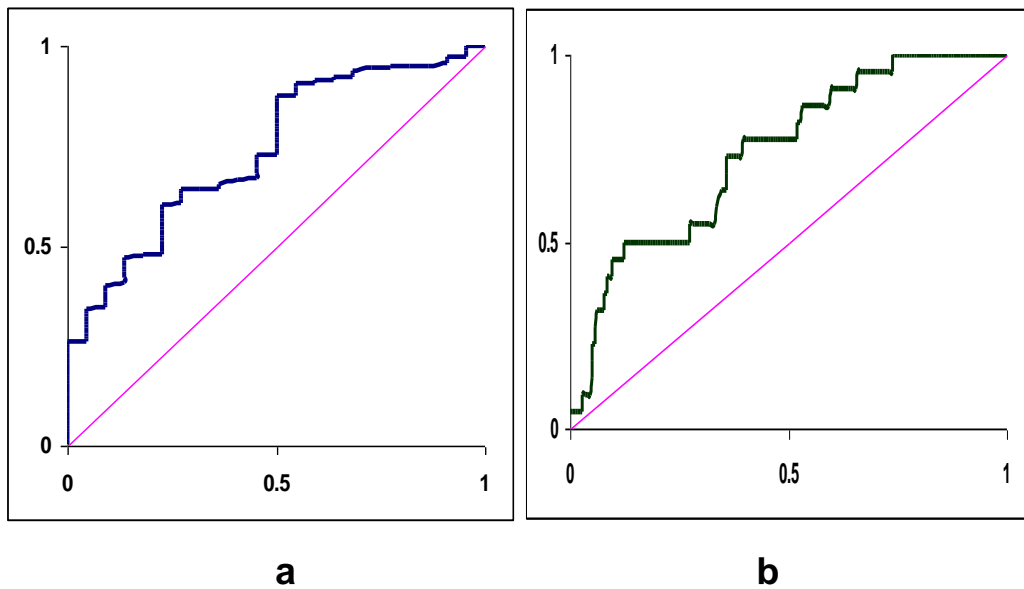


Figure 4-12 OS prediction performance: ROC curves
a: less than five years; b: more than five years

4.4 Categorization of the continues data

There are some benefits considered for the categorization of continues data (Liu et al, 2002; Krzysztof et al., 2007). Some of the data mining methods can only handle categorized data. Furthermore if the number of continuous values of variables (features) is large, it is hard or inefficient to build a model for such data (Krzysztof et al., 2007). On the other hand, it has been argued that categorization leads the model to loss of power, precision and accuracy (Royston et al., 2006; Fedorov et al., 2009). Therefore, in order to assess the optimal method for analysing continuous data, the two parameters with continuous data were chosen: CA125 level and Age. Based on author's supervisor experience and suggestions, CA125 parameter was then categorised into five categories: Less than 35, Between 35 and 200, Between 200 and 600, Between 600 and 1200 and More than 1200. Age parameter was categorized into four categories based upon its quartiles. The dataset was then analysed twice, firstly using all categorised data, secondly using all four parameters as continuous data. Table 4-4 summarizes the prediction performance for categorized and continuous data for six different survival categories.

The prediction performance results for both models are similar. However, accuracy of the model when using continuous data is higher than the model that uses categorized data (Table 4-4). The MAE and RMSE values for the model that uses continuous data is lower compared to the model that uses categorized data (table 4-4).

The possible problems and benefits of categorizing continuous data were discussed at the start of this section. On the other hand, all the models that were used in this research can handle continuous data and the values of continuous variables (age and CA125) are relatively small. Furthermore, the accuracy and the power of the models are the most important factors for this research. The prediction performance results for the model that uses continuous data are better, compared to the model that uses categorized data. Keeping these in mind, the performance of the models were optimal with all two parameters analysed as continuous data.

Number of survival categories	Categorized data			Continuous data		
	Accuracy	MAE	RMSE	Accuracy	MAE	RMSE
2	91.03	0.21	0.32	93.64	0.18	0.24
4	50.16	0.28	0.41	52.37	0.22	0.31
8	40.15	0.18	0.33	42.55	0.13	0.27
16	19.94	0.1	0.26	21.21	0.05	0.26
32	14.32	0.05	0.19	19.95	0.05	0.19
64	4.98	0.03	0.13	9.21	0.03	0.13

Table 4-4 Prediction performance: Categorized data versus Continuous data

4.5 Model analysis by increasing the number of data

The process of data analysis in this research was discussed previously (section 3.5). Briefly at the start of the process 190 cases were randomly selected and initial analyses were performed. In each interval 18 new cases were added to the dataset and the prediction performances were recorded. These processes were continued until all cases were added and analyzed. According to Marzban (2009) and Lee (2010), the increase in number of data can improve the performance of the models. Therefore the increases in the performance of the models were expected.

Figure 4-13 illustrates the effects of increasing the number of cases on the accuracy of the model for all survival categories. As discussed, at the start of the process 190 cases were randomly selected. At each interval 18 new cases were added to the dataset and the performance of the model was analyzed and recorded. As expected for all survival categories the accuracy of the model slightly improved as the number of cases increased. The only exception was when the number of cases reached 280. At this point the accuracy of the model was slightly reduced compared to the previous dataset (262 cases). This reduction in accuracy may be caused by completely new values that the latest number of cases were held.

As expected the performance of the model improved as the number of cases increased. The performance of the model continued to increase until all available cases were added to the dataset. Therefore it is impossible to discover the training stop point (as discussed in section 3.4). As a result, adding new cases to the dataset may improve the performance of the models.

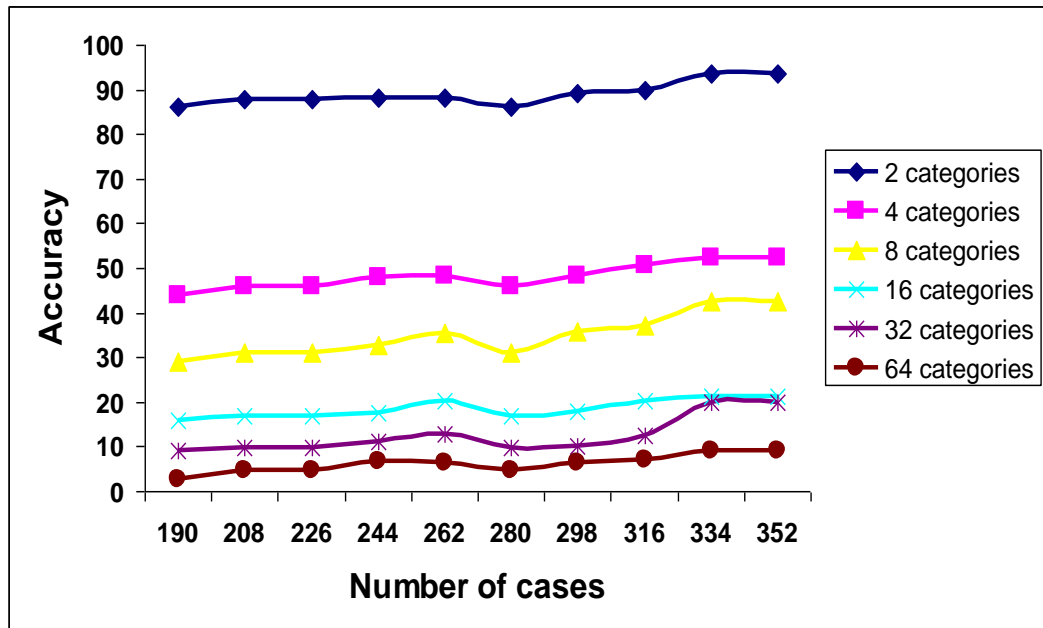


Figure 4-13 Effects of increasing number of cases on accuracy of the model
190 cases randomly selected at the start of the process. At each interval 18 new cases were added to the dataset and the performance of the model was analysed and recorded.

4.6 The best model

Previously (section 1.3.1) three different methods were introduced for creating models. These models are Artificial Neural Network (ANN), Bayesian Network (BN) and Decision Tree (DT). These three models were assessed to discover the best method for this research.

Figure 4-14 illustrates the comparison results of ANN, NB and DT methods. Figure 4-14 a, demonstrates the accuracy of predictions of the above methods. ANN in 78% of all analyses achieved the higher accuracy compared to two other methods. DT gained better accuracy results of 13% compared to BN of 9%. Figure 4-14 b, demonstrates the MAE results for the above models. In 96% of all analyses ANN achieved the lower values of MAE. DT placed in second (3%) and BN in last place (1%). Figure 4-14 c, demonstrates the RMSE results. In 87% of all analyses, ANN achieved lower RMSE results. DT was in second place (8%) and BN came last (5%). These results demonstrate that ANN is the optimum prediction method which is well suited to the dataset available for this research.

In order to achieve better comparison, the ROC curve for all the above were produced (for survival less/more than median value). Figure 4-15 illustrates the ROC curves for ANN, DT and BN where the number of survival categories was two. Figure 4-15 a, was produced using the ROC analysis for survival less than the median value. This analysis indicated that ANN (0.72) performed better than the other two methods (DT: 0.57; BN: 0.51). The area under ROC curve produced by ANN was more than the areas produced by the other two methods. Figure 4-15 b, demonstrates the results of ROC analysis when the survival rate was more than the median value. The area under ROC curve for ANN results (0.72) was more than the areas produced by the other two methods (DT: 0.60; BN: 0.55).

In most cases ANN achieved the highest accuracy results (78% of all analyses). ANN also in most number of cases (96% of all analyses) gained the lower value of MAE. In 87% of all analyses ANN scored the lower values of RMSE. The results of ROC analysis (when the number of survival categories was two), indicated that the area under ROC curve that produced by ANN was more than the areas produced by other two methods.

Based on the results of accuracy, of MAE and RMSE (Figure 4-14), it can be concluded that the best was ANN.

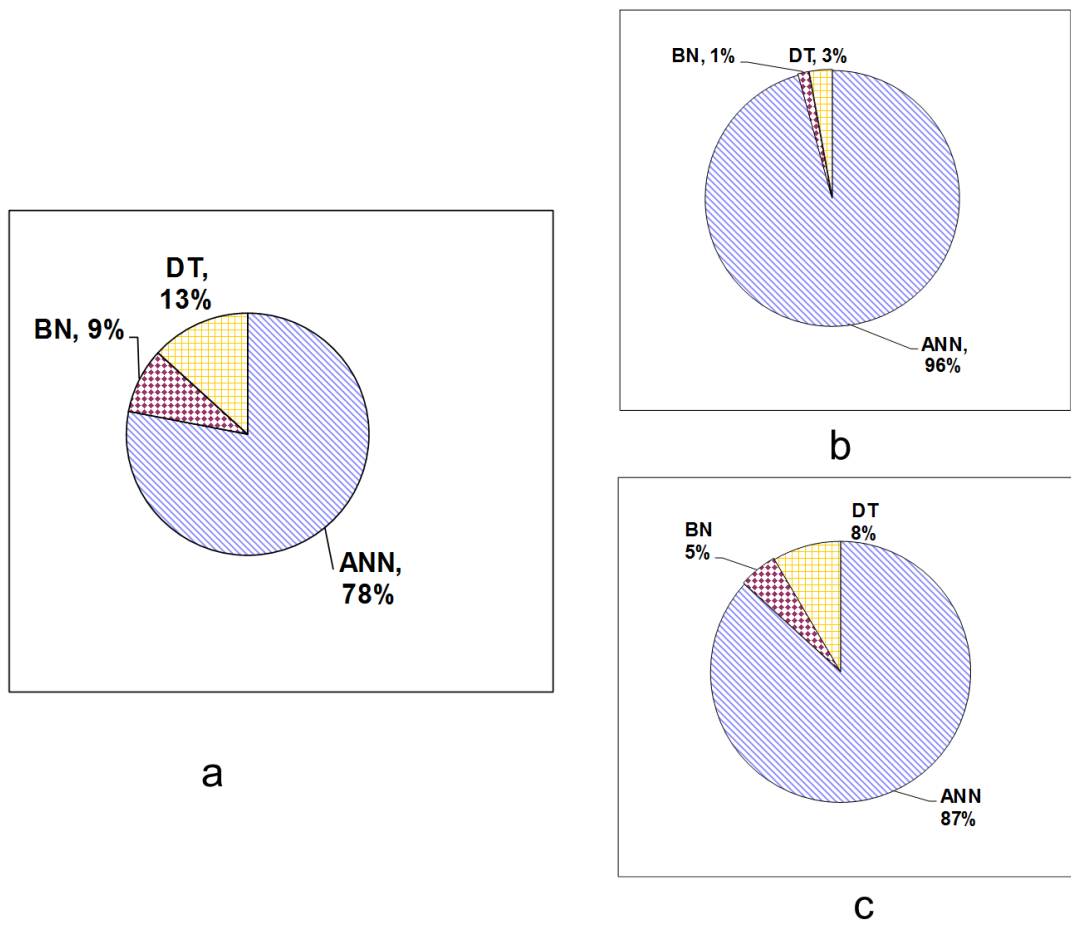


Figure 4-14 Prediction performance comparison (ANN, DT and BN)
a: Accuracy, b: MAE and c: RMSE

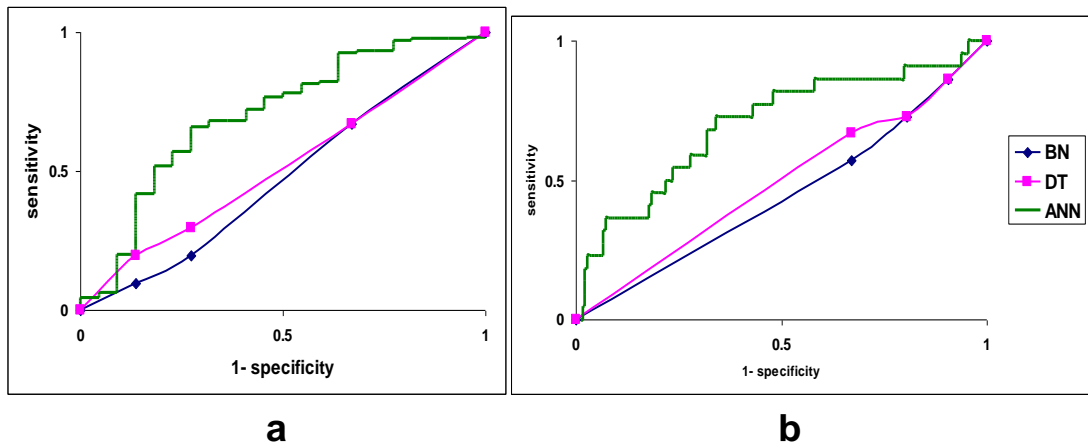


Figure 4-15 Model comparison using ROC curve
(a): survival is less than median value. (b): survival is more than median value

4.7 Comparison with conventional statistics

In order to have a complete comparison between models, Logistic Regression models as standard conventional statistical prediction models were developed. The results of Logistic Regression Models were then compared against the ANN results.

Figure 4-16 compares the prediction performance results of ANN and Logistic Regression. Figure 4-16 a, indicate that in 75% of all analysis ANN achieved higher accuracy results compared to Logistic Regression (25%). Figure 4-16 b, demonstrate the results of MAE for ANN and Logistic Regression. In 98% of all analyses, ANN produced lower values of MAE compared to Logistic Regression (2%). Figure 4-16 c, demonstrate the results of RMSE for ANN and Logistic Regression. In 92% of all analyses ANN scored lower values of RMSE compared to Logistic Regression (8%).

Figure 4-17 compares the results of ROC analysis of ANN and LR when the number of survival categories was two. Figure 4-17 a, represents the results of ROC analysis when the survival was less than the median. These results indicated that the area under the ROC curve for ANN (0.72) was more than the area under the ROC curve for LR (0.62). Figure 4-17 b, represents the results of ROC analysis when the survival was more than the median. These results indicated that the area under the ROC curve for ANN (0.72) was more than the area under the ROC curve for LR (0.62).

Based on total of analyses performed, in 75% of the cases ANN achieved higher results of accuracy. Meanwhile, ANN in most number of analyses (98%) achieved the lower values of MAE and in 92% of all analyses scored lower values of RMSE. Also the *area under the ROC curve* when the number of survival categories was two, for ANN was 0.72 and for LR was 0.62.

Based on the results of Accuracy, MAE and RMSE, it can be concluded that ANN outperforms Logistic Regression.

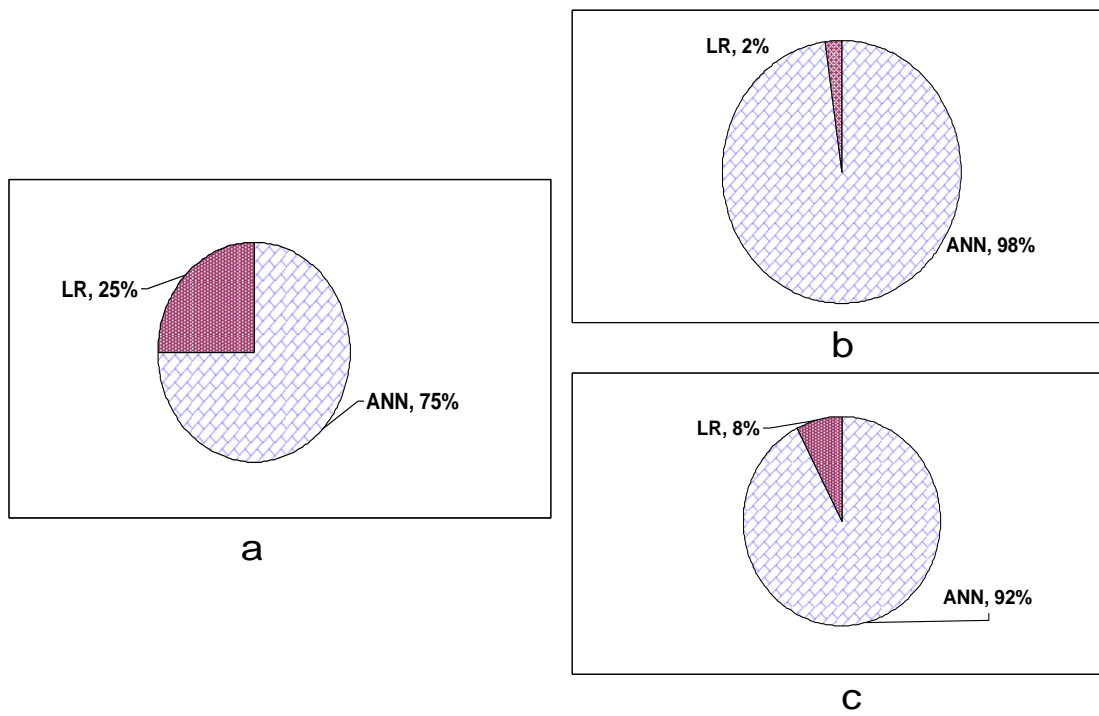


Figure 4-16 ANN vs. Logistic Regression
a: Accuracy, b: MAE and c: RMSE

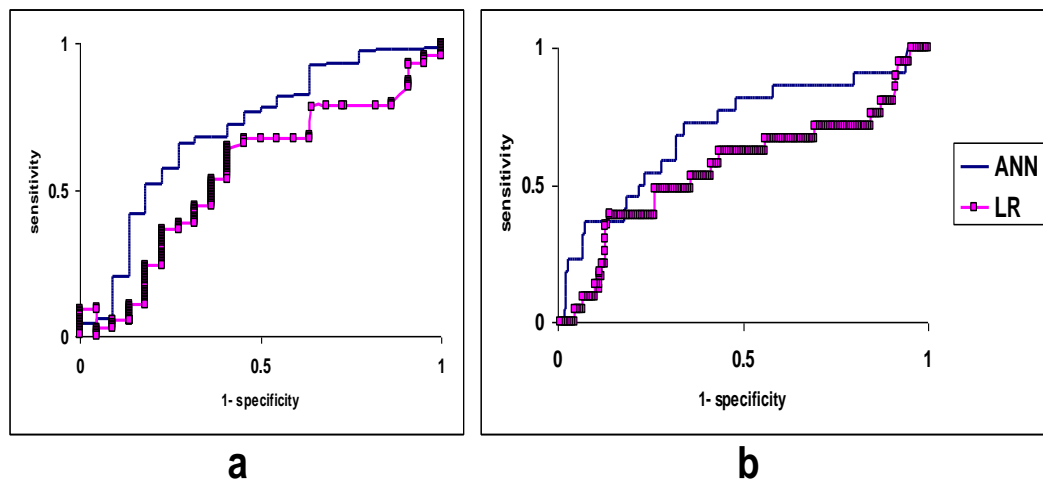


Figure 4-17 ANN vs. Logistic Regression (ROC curves)
(a): survival is less than the median value. (b): survival is more than the median value

4.8 Comparison with pervious work

Recently there were some attempts to predict the overall survival of an ovarian cancer patient (Teramukai et al., 2007; Chi et al., 2008; Gerestein et al., 2009). Gerestein et al. (2009) developed a 'Nomogram' which predicts five year survival rates of patient based on five parameters. This Nomogram calculates the five year survival rate to a percentage probability. This research had c statistic of 0.67.

As discussed previously (Section 3.3.2) using ANN, a model was developed to predict the five year survival of an ovarian cancer patient. ANN produced a good prediction model with an accuracy of 92.7%. The MAE of this model was 0.16 and RMSE was 0.30. The AUC for this model was 0.74. Based on the AUC (which is same as c statistic), it can be concluded that the model that developed using ANN, outperformed the previously developed model. The prediction range of the present model (2, 4, 8, 16, 32 and 64 categories) is much wider than the previously developed model (2 categories).

The other possible advantage of the current model is that as discussed previously (Section 4.5), the prediction results can be improved by increasing the number of available data.

It also can be argued the present model is more user friendly. The user only needs to enter available variables and the system would show the prediction results.

4.9 Summary of the chapter

The analyses were performed to discover the optimum size of the feature set that maximizes the prediction accuracy. Based on the results of accuracy, of MAE and RMSE, it can be concluded that the combination of three features was the optimum feature set. Later on, all possible permutations and combinations of the features were discussed and for all possible feature set sizes the best possible combination(s) was analyzed. Briefly, Outcome of the surgery is the most important feature if a single feature dataset was selected. The combination of Outcome of the surgery and CA125 is the most important feature set, if a two features dataset was selected. The combination of FIGO, Outcome of the surgery and CA125 is the most important feature set, if a three features dataset was selected. The combination of Age, Grade, Outcome of the surgery and CA125 is the most important feature set, if a four features dataset was selected. The combination of Age, Grade, Histology, Outcome of the surgery and CA125 is the most important feature set, if a five features dataset was selected. The above results were also

compared against the results of the most important features that were identified using feature selection techniques. All of the techniques failed to discover the optimum feature set; however GA and TPP performed better compared to two other techniques.

The survival was divided into 64 categories based on the median value available in the dataset. The prediction performance for all categories were recorded and analyzed. In summary the prediction performance of the model decreased as the number of categories increased. The prediction accuracy when the number of categories was two (less or more than median) was around 90%. The best model for this number of categories produced 93.64% of prediction accuracy. The prediction recorded as 9.21% for 64 survival categories. The AUC for the prediction of two survival categories was 0.72. The AUC for prediction of four survival categories was around 0.61.

The potential advantages and disadvantages of categorization of continuous data were discussed. The Age and CA125 as two continuous features were categorized and the prediction performance results were compared to the prediction performance results of uncategorized features. Based on the results of accuracy, of MAE, RMSE and ROC values, this research concluded that analysis using continuous data was superior to categorical data. The five year survival rate (OS) of the patients was also analysed in this chapter. In summary ANN produced a good prediction model by accuracy of 92.7%. The MAE of this model was 0.16 and RMSE was 0.30. The AUC for this model was 0.74.

The effects of increasing number of cases were also analysed and discussed. In summary, by increasing the number of cases the prediction performance of the models were improved.

Considering all the performed analyses, this research concluded that ANN was the best analysed model. In summary, these results were based on number of times ANN achieved the higher values of Accuracy and lower values of MAE and RMSE. ANN in 78% of all analyses achieved the higher values of accuracy. ANN also in 96% of all analyses achieved the lower values of MAE. ANN in 87% of all analyses produced the lower values of RMSE. The AUC produced by ANN for two survival categories was 0.72 compared to DT of 0.60 and BN of 0.55.

The prediction performance results of ANN as the best analysed model was compared to the prediction performance results of Linear Regression. This research concluded that ANN outperformed Linear Regression. In summary in 75% of all analyses ANN

achieved higher accuracy results. ANN also in 98% of all analyses scored lower values of MAE. Furthermore, ANN in 92% of all analyses scored lower values of RMSE. The AUC produced by ANN for two survival categories was 0.72 compared to LR of 0.62.

The previous survival prediction models that had been produced by other researchers were briefly discussed. The results of other methods were discussed and compared to the present model. In summary, the present model performed better in comparison to the previously developed models, it may be easier to use and the prediction results can be improved as the number of cases increases.

Chapter 5 Surgical Prediction

5.1 Introduction

This chapter investigates the potential benefits of predicting the outcome of surgery. The possibility of developing models for such a prediction was investigated and analysed. Artificial intelligence and statistical models were developed for this investigation. The complete comparison of these two techniques is made available in this chapter. This chapter also investigates and introduces the most important variables (factors) in the dataset for predicting the outcome of the surgery.

5.2 Statement of the problem

The benefits of primary surgery for ovarian cancer patients were discussed previously (section 1.2.1.8). Despite aggressive attempts at optimal cytoreduction, many patients are left with sub-optimal disease after surgery. As discussed previously (section 1.2.1.10), the current treatment of ovarian cancer is chemotherapy after surgery, however research (Vergote et al., 2005) indicates that the outcome for patients left with sub-optimal disease after surgery is worse than that of optimally debulked patients. In fact a consistent finding is that residual bulk of disease following surgery is the greatest prognostic factor in ovarian cancer (Edmondson et al., 2008). Indeed it has now been suggested (Stack and Fishman, 2009) that the aim of surgery should be to remove the entire cancer from the patient. A pre-operative assessment tool, capable of predicting surgical outcome is therefore needed to identify cases where complete cytoreduction is achievable or cases with early stages of cancer.

In recent years a new possible treatment is recommended for cases where complete or optimal cytoreduction appears unlikely. The recommended treatment is chemotherapy prior to surgery (Neoadjuvant chemotherapy). Neoadjuvant chemotherapy seems to be the best solution for some patients as it may reduce the size of the tumour and boost the possibility of the complete/optimal cytoreduction during surgery (Stack and Fishman, 2009). Neoadjuvant chemotherapy appears to be beneficial in the following situations (Stack and Fishman, 2009):

- Where complete or optimal cytoreduction appears unlikely due to the position of the tumour.
- Where the patient is too weak for surgery. These types of patients may benefit from chemotherapy as the size of the tumour will reduce with relief of symptoms after few cycles of chemotherapy.

- Where the patient suffers from poor nutritional status. Neoadjuvant chemotherapy may reduce the symptoms of cancer and allow the patient to improve their nutritional status. Poor nutritional intake is related to poor outcome after surgery including delayed wound healing (Windsor et al., 1998),
- Where the patient suffers from depression. A study on cervical cancer patients (Plante et al., 2006), considered depression as one of the factors causing more difficulty for patients to be mobile and more active. Neoadjuvant chemotherapy may reduce the effects of the disease on patients, improve their mood and help them to recover faster.
- Where the patient lives in a country where the surgery admission process takes a long time. The long waiting lists for surgery may decrease the chances of survival for patients, as most cases of ovarian cancer present in the late stages of the cancer (section 1.2.1.1).

The other possible benefit of neoadjuvant chemotherapy is that it gives an early indication of tumour response to chemotherapy (Stack and Fishman, 2009). If the tumour did not respond to neoadjuvant chemotherapy, the chemotherapy can be changed for further treatment.

At present, there is no strong evidence to support the role of neoadjuvant chemotherapy (Stack and Fishman, 2009). However, research (Vergote et al., 2005) suggests higher rates of optimal cytoreduction after neoadjuvant chemotherapy. On the other hand, the same study concluded similar outcomes for patients with primary surgery and neoadjuvant chemotherapy. Other research (Bristow and Chi, 2006) suggested that the overall survival rates of patients did not improve using the neoadjuvant approach; on the other hand this method was recommended for patients with difficulties as discussed previously. More recently, research (Vergote et al., 2010) did not conclude the replacement of Neoadjuvant treatment by primary surgery. The complete results of two large randomised studies, the Chemotherapy or Upfront Surgery (CHORUS) and European Organization for Research and Treatment of Cancer (EORTC) 55971, are not published yet. These two large studies may resolve the issues surrounding the neoadjuvant treatment.

As discussed, the prediction of the outcome of the surgery may be very beneficial to the patients. This prediction may help clinicians to suggest Neoadjuvant treatment for patients with possible outcome of sub-optimal cytoreduction after surgery. Allen (2010) argues that a clinical model that can predict which patients will undergo optimal

cytoreduction would be very useful and he emphasizes that “at present, no such model exists”. In the past, Bristow et al. (2000) and Funt et al. (2004) developed the models for predicting outcome of the surgery using the findings on CT scans. These models were either very complex to assess or they did not conclude any significant results.

5.3 Residual prediction

As there is no model available to assist clinicians with such predictions we considered the potential of Machine learning and artificial intelligence models (section 1.3.1). This section investigates the opportunities of employing such models for predicting the outcome of surgery. The factors in the datasets available for this investigation (section 2.2), can all be collected prior to surgery. Therefore using the available datasets (section 2.2), and employing machine learning and artificial intelligence model, this section investigates the performance of prediction of such models.

Prior to this investigation it is essential to make some adjustments in the dataset as stated below:

- The prediction goal (class) has to be changed from the survival rates to outcome of the surgery. In the main datasets, the survival rates of the patients are set as the prediction goal of the models, as we are going to predict the outcome of the surgery the prediction goal has to be changed to outcome of the surgery.
- The survival rate has to be omitted from the dataset. Prior to the surgery the survival rates of the patients are not known, therefore it cannot be used as a prediction factor.
- The CA125 is a continuous data (section 2.2), categorization of these values will be performed to investigate the categorization of the performance of the models. Then the results will be compared with the performance of the models using actual values of CA125. The categorization criteria are as stated below:
 1. CA125 greater than 0 and less than 35.
 2. CA125 greater than 35 and less than 200.
 3. CA125 greater than 200 and less than 600.
 4. CA125 greater than 600 and less than 1200.
 5. CA125 greater than 1200.
- The possible outcomes of the prediction are: complete, optimal and sub-optimal cytoreduction (section 1.2.1.8). However as discussed previously (section 5.2) it

would be beneficial to distinguish between the following categories in cytoreduction outcome:

1. The possibility of complete cytoreduction versus optimal and sub-optimal cytoreduction.
2. The possibility of complete and optimal cytoreduction versus sub-optimal cytoreduction.
3. The possibility of complete cytoreduction versus optimal cytoreduction versus sub-optimal cytoreduction.

In order to make a comparison with conventional statistical methods, regression models were built using SPSS and the prediction performances of these models are made available.

5.3.1 Residual prediction: Complete/Optimal/Sub-optimal cytoreduction

Using ANN (section 1.3.1.2) the possibility of predicting the outcome of surgery for all possible outputs will be investigated. First the model was trained and evaluated using the actual values of CA125, then CA125 is categorized to the mentioned categories and the model was trained and evaluated. The evaluation process is similar to the process that defined previously (section 3.5) and 10 fold cross validation (section 3.2.3) is employed.

Table 5-1 summarizes the results of the prediction performance for **actual values** (no categorization) of CA125 for the best model produced using **ANN**. The accuracy of the model was 59.75% and the MAE value was 0.34 and RMSE value was 0.42. The AUC for complete cytoreduction curve was 0.75. The AUC for optimal cytoreduction was 0.66 and the AUC for suboptimal cytoreduction was 0.71.

Table 5-2 summarizes the results of the prediction performance for **actual values** (no categorization) of CA125 for the best model produced using **Logistic Regression**. The MAE value for this model was 0.61 and RMSE value was 0.75. The AUC for complete cytoreduction was 0.24. The AUC for optimal cytoreduction was 0.70 and for sub-optimal was 0.60.

Figure 5-1 illustrates the ROC curves for **actual values** (no categorization) of CA125 for the best model produced using **ANN** (complete cytoreduction: 0.75. optimal cytoreduction: 0.66 and suboptimal cytoreduction: 0.71).

Figure 5-2 illustrates the ROC curves for **actual values** (no categorization) of CA125 for the best model produced using **Logistic Regression** (complete cytoreduction: 0.24, optimal cytoreduction: 0.70 and sub-optimal: 0.60).

ANN produced two good models based on the curve analysis (Figure 5-1) for complete cytoreduction (0.7587) and suboptimal cytoreduction (0.7123), however the results for optimal cytoreduction (0.6609) prediction is fair. These results are much higher than the prediction performance produced by logistic regression (Figure 5-2). The error rates of prediction using ANN (Table 5-1) are much lower than logistic regression (Table 5-2).

Measurement		Result
Accuracy		59.75%
Mean absolute error		0.34
Root mean squared error		0.42

Outcome	Recall	Precision	AUC
Complete	0.636	0.8	0.7587
Optimal	0.691	0.428	0.6609
Sub-optimal	0.28	0.477	0.7123

Table 5-1 Prediction performance of results for complete/optimal/sub-optimal cytoreduction when CA125 is not categorized using ANN

Measurement	Result
Mean absolute error	0.61
Root mean squared error	0.75

Outcome	AUC
Complete	0.2440
Optimal	0.7060
Suboptimal	0.6000

Table 5-2 Prediction performance results for complete/optimal/suboptimal cytoreduction when CA125 is not categorized using Logistic Regression

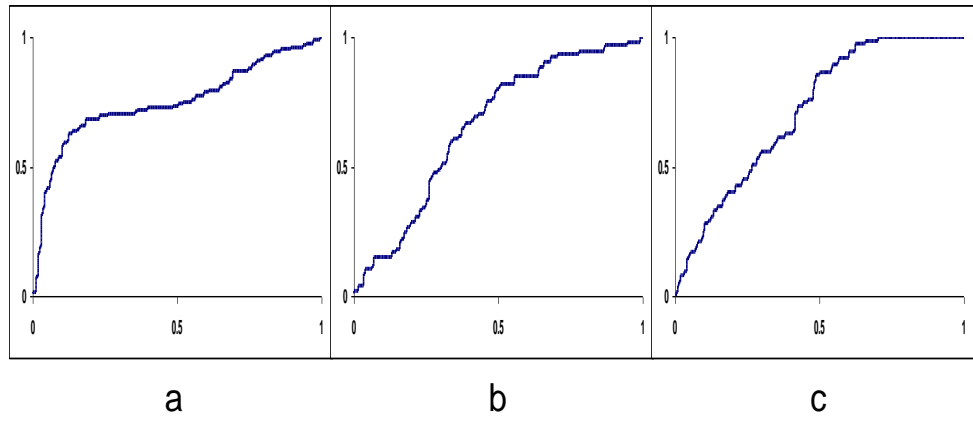


Figure 5-1 ROC curve (ANN): Complete / Optimal /Sub-optimal cytoresduction (CA125 not categorized)
 (a) complete cytoresduction(0.75);(b) optimal cytoresduction(0.66); (c) sub-optimal cytoresduction(0.71)

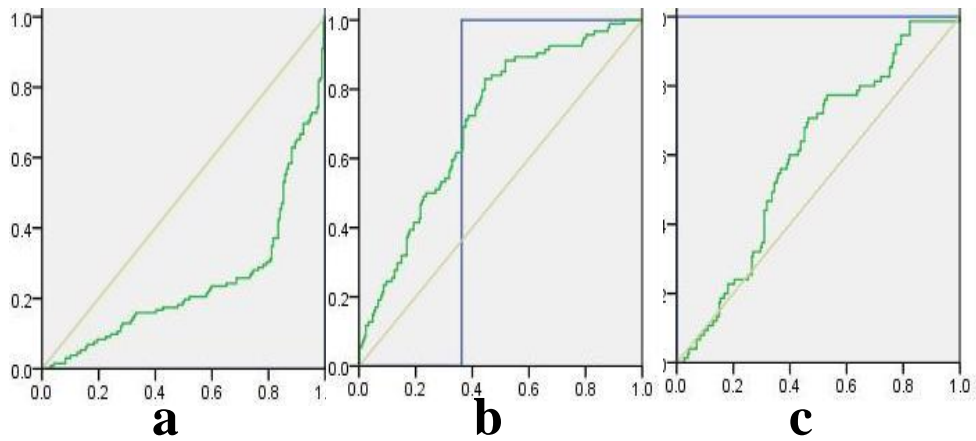


Figure 5-2 ROC Curve (Logistic Regression): Complete / Optimal /Suboptimal cytoresduction (CA125 not categorized)
 (a) complete cytoresduction(0.24);(b) optimal cytoresduction(0.70); (c) suboptimal cytoresduction(0.60)

Table 5-3 summarizes the results of the prediction performance for **categorized values** of CA125 for the best model produced using **ANN**. The accuracy of this model was 53.15%. The MAE value was 0.35 and the RMSE value was 0.43. The AUC for complete cytoreduction was 0.7253. The AUC for optimal cytoreduction was 0.6416 and AUC for sub-optimal cytoreduction was 0.7174.

Table 5-4 summarizes the results of the prediction performance for **categorized values** of CA125 for the best model produced using **Logistic Regression**. The MAE value for this model was 0.61 and RMSE value was 0.75. The AUC for complete cytoreduction was 0.23. The AUC for optimal cytoreduction was 0.72 and AUC for sub-optimal cytoreduction was 0.59.

Figure 5-3 illustrates the ROC curves for **categorized values** of CA125 for the best model produced using **ANN** (complete cytoreduction: 0.7253. optimal cytoreduction: 0.6416 sub-optimal cytoreduction: 0.7174).

Figure 5-4 illustrates the ROC curves for **categorized values** of CA125 for the best model produced using **Logistic Regression** (complete cytoreduction: 0.23. optimal cytoreduction: 0.72 sub-optimal cytoreduction: 0.59).

ANN produced two good models based on the curve analysis (Figure 5-3) for complete cytoreduction (0.7253) and suboptimal cytoreduction (0.7174), however the results for optimal cytoreduction (0.6416) prediction is fair. These results are much higher than the prediction performance produced by logistic regression (Figure 5-4). The error rates of prediction using ANN (Table 5-3) are much lower than logistic regression (Table 5-4).

Discussion: ANN produced the same accuracy and error rates for actual values of CA125 compared to categorized values of CA125. Also there are no noticeable differences in recall, precision and AUC values. Logistic regression also produced the same results when using actual values of CA125 compared to categorized values of CA125.

Conclusion: Considering the prediction performance results produced by ANN (Table 5-1 and Table 5-3) and the prediction performance results of Logistic Regression (Table 5-2 and Table 5-4), it can be concluded that ANN outperforms Logistic Regression. Also it can be concluded that the categorization of CA125 did not improve the prediction performance results.

Measurement		Result	
Accuracy		53.15%	
Mean absolute error		0.35	
Root mean squared error		0.43	

Outcome	Recall	Precision	AUC
Complete	0.659	0.784	0.7253
Optimal	0.681	0.438	0.6416
Sub-optimal	0.253	0.432	0.7174

Table 5-3 Prediction performance results for complete/optimal/sub-optimal cytoreduction when CA125 is categorized using ANN

Measurement	Result
Mean absolute error	0.61
Root mean squared error	0.75

Outcome	AUC
Complete	0.2320
Optimal	0.7220
Sub-optimal	0.5980

Table 5-4 Prediction performance results for complete/optimal/sub-optimal cytoreduction when CA125 is categorized using Logistic Regression

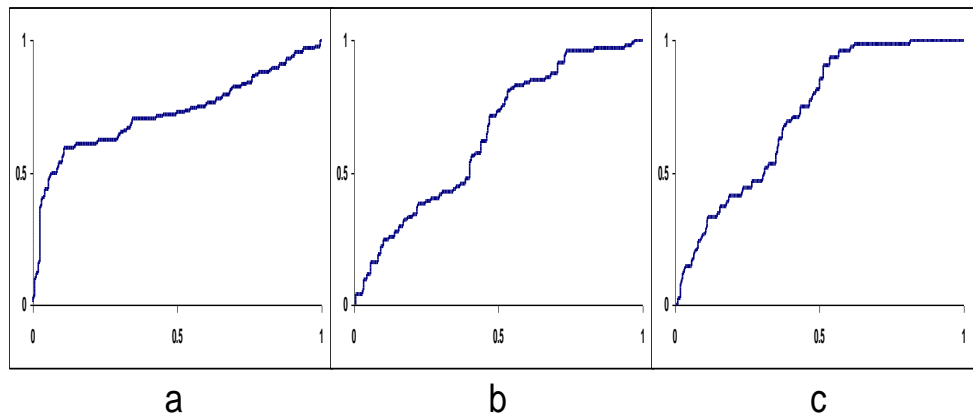


Figure 5-3 ROC curve (ANN): Complete / Optimal /Suboptimal cytoresduction (CA125 categorized)
(a) complete cytoresduction(0.72);(b) optimal cytoresduction(0.64); (c) suboptimal cytoresduction(0.71)

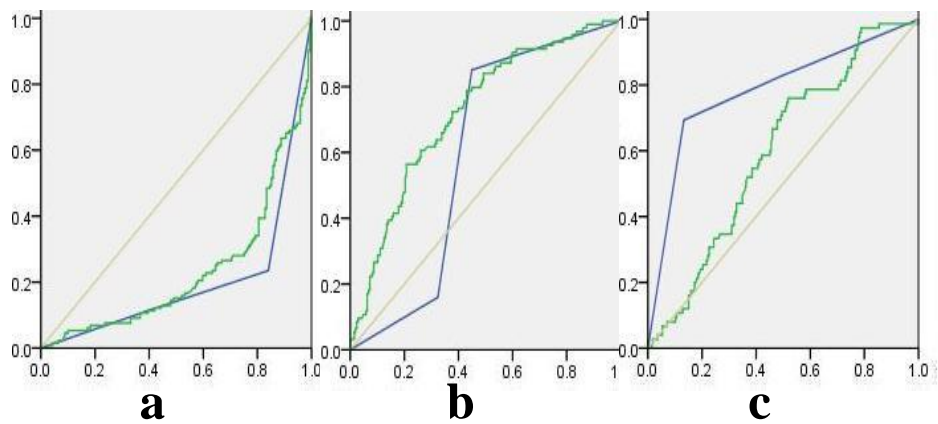


Figure 5-4 ROC Curve (Logistic Regression): Complete / Optimal /Suboptimal cytoresduction (CA125 categorized)
(a) complete cytoresduction(0.23);(b) optimal cytoresduction(0.72); (c) suboptimal cytoresduction(0.59)

5.3.2 Residual prediction: Complete and Optimal versus Suboptimal cytoreduction

At the start of this process, the dataset was redefined, so the outcome of the surgery is changed to either ‘complete or optimal’ or ‘suboptimal’. Similar to previous analysis (Section 5.3.1), first the model is trained and evaluated using categorized values of CA125, then the investigation is repeated using the actual values of CA125. Similar to the process of previous section, the 10 fold cross validation is used for the evaluation process.

Table 5-5 summarizes the results of the prediction performance for **actual values** (no categorization) of CA125 for the best model produced using **ANN**. The accuracy of this model was 77.75%. The MAE value was 0.29 and the RMSE value was 0.38. The AUC for complete or optimal cytoreduction was 0.7310 and the AUC for sub-optimal cytoreduction was 0.7310.

Table 5.6 summarizes the results of the prediction performance for **actual values** (no categorization) of CA125 for the best model produced using **Logistic Regression**. The MAE for this model was 0.37 and the RMSE was 0.43. The AUC for the complete or optimal cytoreduction was 0.3970 and AUC for sub-optimal cytoreduction was 0.6330.

Figure 5-5 illustrates the ROC curves for **actual values** (no categorization) of CA125 for the best model produced using **ANN** (complete or optimal cytoreduction: 0.73, sub-optimal cytoreduction: 0.73)

Figure 5-6 illustrates the ROC curves for **actual values** (no categorization) of CA125 for the best model produced using **Logistic Regression** (complete or optimal cytoreduction: 0.39, sub-optimal cytoreduction: 0.63)

ANN produced two good models based on the AUC values (Table 5-5). The AUC results for complete or optimal cytoreduction (0.7310) and for suboptimal cytoreduction (0.7310) indicate a very good prediction performance. The error rates are relatively small and the accuracy (77.40 %) is good. On the other hand the AUC results of Logistic Regression model are 0.3970 for complete or optimal cytoreduction and 0.6330 for suboptimal cytoreduction (Table 5-6). The error rates produced by Logistic Regression are higher compared to ANN.

Measurement		Result
Accuracy		77.75 %
Mean absolute error		0.29
Root mean squared error		0.38

Outcome	Recall	Precision	AUC
Complete or Optimal	0.359	0.775	0.7310
Sub-optimal	0.147	0.647	0.7310

Table 5-5 Prediction performance results for (complete or optimal) cytoreduction versus sub-optimal cytoreduction when CA125 is not categorized using ANN

Measurement	Result
Mean absolute error	0.37
Root mean squared error	0.43

Outcome	AUC
Complete or Optimal	0.3970
Sub-optimal	0.6030

Table 5-6 Prediction performance results for (complete or optimal) cytoreduction versus sub-optimal cytoreduction when CA125 is not categorized using Logistic Regression

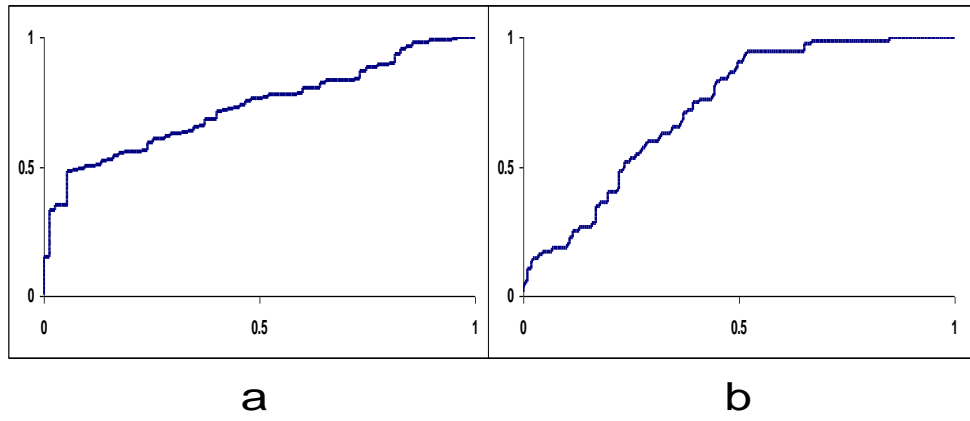


Figure 5-5 ROC curve (ANN): (Complete or Optimal) cytoreduction versus Sub-optimal cytoreduction (CA125 not categorized)
(a) complete or optimal cytoreduction(0.73);(b) sub-optimal cytoreduction(0.73);

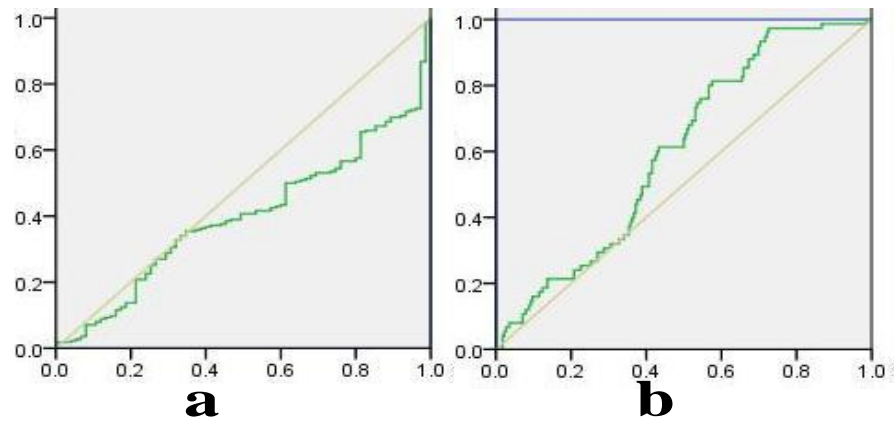


Figure 5-6 ROC curve (Logistic Regression): (Complete or Optimal) cytoreduction versus Sub-optimal cytoreduction (CA125 not categorized)
(a) complete or optimal cytoreduction(0.39);(b) sub-optimal cytoreduction(0.60);

Table 5-7 summarizes the results of the prediction performance for **categorized values** of CA125 for the best model produced using **ANN**. The accuracy of this model was 75.75%. The MAE value was 0.34 and the RMSE value was 0.31. The AUC for complete or optimal cytoreduction was 0.6290 and the AUC for sub-optimal cytoreduction was 0.6290.

Table 5-8 summarizes the results of the prediction performance for **categorized values** of CA125 for the best model produced using **Logistic Regression**. The MAE value for this model was 0.37 and RMSE value was 0.43. The AUC for complete or optimal cytoreduction was 0.39. The AUC for sub-optimal cytoreduction was 0.60.

Figure 5-7 illustrates the ROC curves for **categorized values** of CA125 for the best model produced using **ANN** (complete or optimal cytoreduction: 0.62, optimal cytoreduction: 0.62).

Figure 5-8 illustrates the ROC curves for **categorized values** of CA125 for the best model produced using **Logistic Regression** (complete or optimal cytoreduction: 0.39, sub-optimal cytoreduction: 0.60).

Based on the prediction performance results of ANN (Table 5-7) two good prediction models were developed. The error rates are small and the accuracy is high (75.75 %). On the other hand the prediction performance results of Logistic Regression are very poor (Table 5-8) and the error rates are higher than ANN.

Discussion: The prediction performance results that were produced using ANN (Table 5-7 and Figure 5-7) and Logistic Regression (Table 5-8 and Figure 5-8) for actual values of CA125 are similar to the results that were produced using categorized values of CA125.

Conclusion: ANN managed to create two good prediction models for the prediction of complete or optimal cytoreduction versus suboptimal cytoreduction. ANN by far outperforms Logistic Regression in this analysis. Also there are no noticeable differences of the results when CA125 is categorized.

Measurement		Result		
Accuracy		75.75 %		
Mean absolute error		0.34		
Root mean squared error		0.31		
Outcome	Recall	Precision	AUC	
Complete or Optimal	0.773	0.751	0.6290	
Sub-optimal	0.027	0.250	0.6290	

Table 5-7 Prediction performance results for (complete or optimal) cytoreduction versus sub-optimal cytoreduction when CA125 is categorized using ANN

Measurement	Result
Mean absolute error	0.37
Root mean squared error	0.43
Outcome	AUC
Complete or Optimal	0.3970
Sub-optimal	0.6020

Table 5-8 Prediction performance results for (complete or optimal) cytoreduction versus suboptimal cytoreduction when CA125 is categorized using Logistic Regression

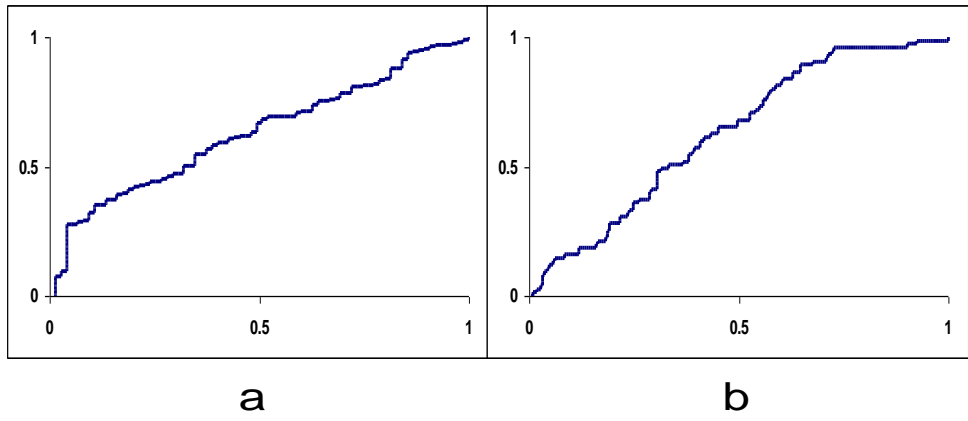


Figure 5-7 ROC curve (ANN): (Complete or Optimal) cytoreduction versus Sub-optimal cytoreduction (CA125 categorized)

(a) complete or optimal cytoreduction(0.62);(b) suboptimal cytoreduction(0.62);

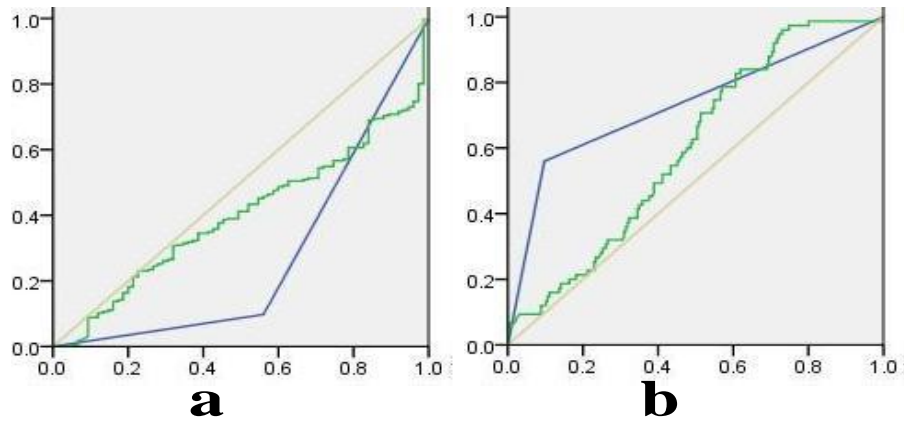


Figure 5-8 ROC curve (Logistic Regression): (Complete or Optimal) cytoreduction versus Sub-optimal cytoreduction (CA125 not categorized)

(a) complete or optimal cytoreduction(0.39);(b) sub-optimal cytoreduction(0.60);

5.3.3 Residual prediction: Complete versus Optimal or Sub-optimal cytoreduction

For this analysis the dataset was redefined in such a way that the outcome of the surgery is either ‘complete cytoreduction’ or ‘optimal or sub-optimal cytoreduction’. The analysis is similar to the previous sections (Section 5.3.1 and Section 5.3.2).

Table 5-9 summarizes the results of the prediction performance for **actual values** (no categorization) of CA125 for the best model produced using **ANN**. The accuracy of this model was 77.74%. The MAE value was 0.32 and the RMSE value was 0.42. The AUC for complete cytoreduction was 0.79 and the AUC for optimal or sub-optimal cytoreduction was 0.80.

Table 5-10 summarizes the results of the prediction performance for **categorized values** of CA125 for the best model produced using **Logistic Regression**. The MAE value for this model was 0.39 and RMSE value was 0.44. The AUC for complete cytoreduction was 0.24 and the AUC for optimal or sub-optimal cytoreduction was 0.69.

Figure 5-9 illustrates the ROC curves for **actual values** (no categorization) of CA125 for the best model produced using **ANN** (complete cytoreduction: 0.79, optimal or sub-optimal cytoreduction: 0.80).

Figure 5-10 illustrates the ROC curves for **actual values** of CA125 for the best model produced using **Logistic Regression** (complete cytoreduction: 0.24, optimal or sub-optimal cytoreduction: 0.69).

ANN managed to produce a very good (AUC: 0.80) and a good (AUC: 0.79) prediction models based on the AUC results (Table 5-9). The Logistic Regression did not manage to produce any worthy models (Table 5-10). Based on the results ANN outperformed Logistic Regression.

Measurement		Result		
Accuracy		77.74 %		
Mean absolute error		0.32		
Root mean squared error		0.42		
Outcome	Recall	Precision	AUC	
Complete	0.621	0.75	0.79	
Optimal or Sub-optimal	0.213	0.76	0.80	

Table 5-9 Prediction performance results for complete cytoreduction versus (optimal or sub-optimal) cytoreduction when CA125 is not categorized using ANN

Measurement	Result
Mean absolute error	0.39
Root mean squared error	0.44
Outcome	AUC
Complete	0.2490
Optimal or Sub-optimal	0.6910

Table 5-10 Prediction performance results for complete cytoreduction versus (optimal or sub-optimal) cytoreduction when CA125 is not categorized using Logistic Regression

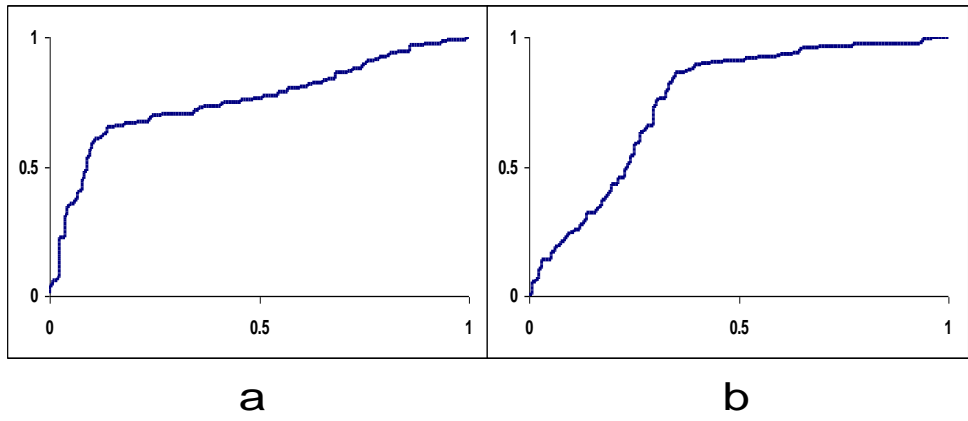


Figure 5-9 ROC curve (ANN): Complete cytoreduction versus (Optimal or Sub-optimal) cytoreduction (CA125 not categorized)
(a) complete cytoreduction(0.79);(b) optimal or sub-optimal cytoreduction(0.80);

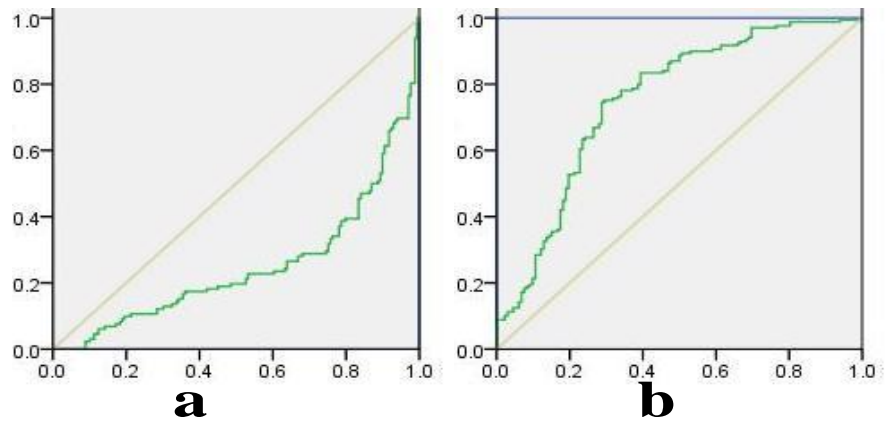


Figure 5-10 ROC curve (Logistic Regression): Complete cytoreduction versus (Optimal or Sub-optimal) cytoreduction (CA125 not categorized)
(a) complete cytoreduction(0.24);(b) optimal or suboptimal cytoreduction(0.69);

Table 5-11 summarizes the results of the prediction performance for **categorized values** of CA125 for the best model produced using **ANN**. The accuracy of this model was 73.75%. The MAE value was 0.34 and the RMSE value was 0.42. The AUC for complete cytoreduction was 0.7710 and the AUC for sub-optimal cytoreduction was 0.7710.

Table 5-12 summarizes the results of the prediction performance for **categorized values** of CA125 for the best model produced using **Logistic Regression**. The MAE value for this model was 0.39 and RMSE value was 0.44. The AUC for complete cytoreduction was 0.24. The AUC for optimal or sub-optimal cytoreduction was 0.76.

Figure 5-11 illustrates the ROC curves for **categorized values** of CA125 for the best model produced using **ANN** (complete cytoreduction: 0.77, optimal or sub-optimal cytoreduction: 0.77).

Figure 5-12 illustrates the ROC curves for **categorized values** of CA125 for the best model produced using **Logistic Regression** (complete cytoreduction: 0.24, optimal or sub-optimal cytoreduction: 0.75).

The prediction performance results of ANN (Table 5-11) indicate two average predictors. However ANN did not manage to produce a worthy model. Similarly Logistic regression only managed to produce one worthy predictor (Table 5-12). The error rates are similar, however on average ANN outperforms Logistic Regression (based on AUC values).

Discussion: Based on the performance prediction results produced by ANN (Table 5-11 and Figure 5-11) and the performance prediction results by Logistic Regression (Table 5-12 and Figure 5-12), categorization of CA125 did not improve the prediction results.

Conclusion: ANN performed better than Logistic Regression. On average, ANN performed better than Logistic Regression. It also can be concluded that categorization of CA125 did not improve the prediction results.

Measurement		Result	
Accuracy		73.75 %	
Mean absolute error		0.34	
Root mean squared error		0.42	
Outcome	Recall	Precision	ROC
Complete	0.644	0.794	0.7710
Optimal or Sub-optimal	0.57	0.758	0.7710

Table 5-11 Prediction performance results for complete cytoreduction versus (optimal or sub-optimal) cytoreduction when CA125 is categorized using ANN

Measurement	Result
Mean absolute error	0.39
Root mean squared error	0.44
Outcome	AUC
Complete	0.2420
Optimal or Sub-optimal	0.7580

Table 5-12 Prediction performance results for complete cytoreduction versus (optimal or sub-optimal) cytoreduction when CA125 is categorized using Logistic Regression

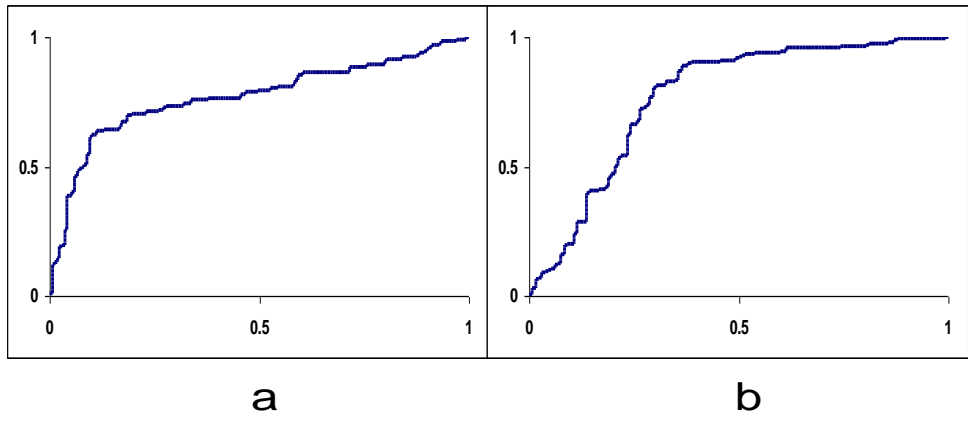


Figure 5-11 ROC curve (ANN): Complete cytoreduction versus (Optimal or Sub-optimal) cytoreduction (CA125 categorized)
 (a) complete cytoreduction(0.77);(b) optimal or suboptimal cytoreduction(0.77);

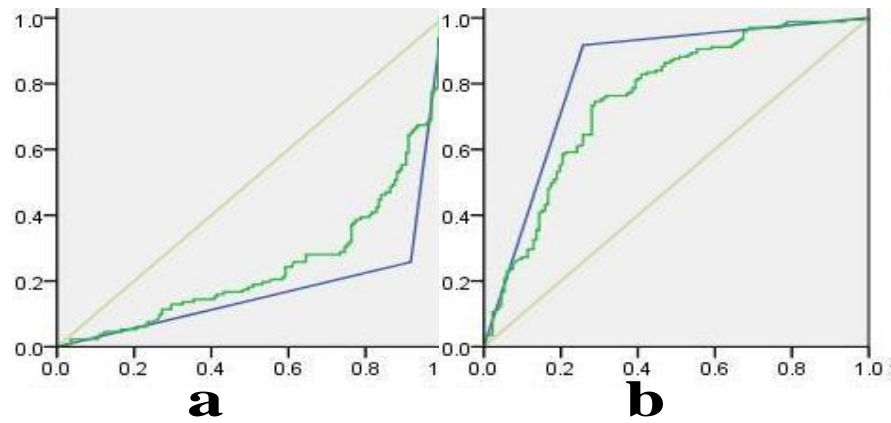


Figure 5-12 ROC curve (Logistic Regression): Complete cytoreduction versus (Optimal or Sub-optimal) cytoreduction (CA125 categorized)
 (a) complete cytoreduction(0.24);(b) optimal or sub-optimal cytoreduction(0.75);

5.4 Discovering important factors

The available dataset was discussed previously. The dataset includes five variables: CA125, Age, Histology type, Grade and FIGO stage and the target variable is the outcome of the surgery. In order to identify the most important factors (variables) the methods of feature selection (section 1.3.2) are deployed. The following table (Table 5-13) summarizes the results of this investigation.

	All three outcomes	Complete vs. Optimal or suboptimal	Complete or Optimal vs. Suboptimal
TPP	G, H, F	F, G	F, H, G
IG	F, G	F, G, H	F, G
GA	F	F, G	F
PCA	F, H, G	F, G	F, G

Table 5-13 The most important variable identified using feature selection methods
(G): Grade, (H): Histology type, (F): FIGO stage

Discussion: The results of using feature selection techniques indicated that the most important factor in the dataset for predicting the outcome of surgery is FIGO stage. Grade of the tumour and Histology are other important factors.

5.5 Summary of the chapter

The benefits of predicting the outcome of the surgery were investigated in this chapter. As discussed, currently there is no such system for predicting the outcome of the surgery. Prediction of the outcome of the surgery has many potential benefits for the patients. Clinicians may offer new treatment pathways (e.g. neoadjuvant chemotherapy) for a patient with high chance of suboptimal cytoreduction. The potential benefits of neoadjuvant chemotherapy were also investigated in this chapter. ANN and Logistic Regression predicting models were developed to predict the outcome of the surgery. Three major investigations were deployed: the prediction of all outcomes (complete or optimal or suboptimal), the prediction of complete or optimal cytoreduction versus suboptimal cytoreduction and finally the prediction of complete cytoreduction versus optimal or suboptimal cytoreduction. CA125 as the only continuous variable in the dataset was categorized. The models that can distinguish between complete or optimal cytoreduction versus suboptimal cytoreduction had the best performance compared to the other two categories. The most important variables (factors) for predicting the

outcome of the surgery were identified. These important variables are: FIGO stage, Grade and Histology.

Based on the results, it can be concluded that ANN outperforms Logistic Regression. Furthermore, categorization of CA125 did not improve the prediction results.

Chapter 6 Gene analysis for classification of Homologous recombination

6.1 Introduction

An increasing understanding of the biology of cancer coupled with the availability of high throughput techniques provides new challenges in cancer research. These techniques allow the generation of large datasets. These datasets may comprise data describing DNA (mutations or epigenetic information), RNA or protein. The interrogation of these datasets is challenging, in part because of the size of the datasets but also because of the signal to noise problems. This chapter introduces the concept of a new treatment for ovarian cancer, the PARP inhibitors, and a DNA repair pathway, homologous recombination (HR) which may be used to stratify therapy. Following this is a description of a dataset generated to investigate this area followed by experiments using the model to investigate the role of artificial intelligence in developing predictive biomarkers. This analysis combined with the models that developed in previous chapters, may introduce the suitable treatment pathway for the management of the patients.

6.2 statement of the problem

DNA is damaged many times during each cell cycle, and this damaged DNA has to be repaired (Karp, 2009). This damage is repaired by DNA repair pathways. One of the DNA repair pathways is homologous recombination (HR) and two of the important proteins in this pathway are BRCA1 and BRCA2 (Yap et al., 2010). There is strong evidence that patients with germline mutations in BRCA1 and BRCA2 have a high risk of ovarian cancer (Woosetr and Weber, 2003). Poly (ADP-ribose) polymerase (PARP) is an important protein in another DNA repair pathway, called base excision repair. Inhibiting PARP results in deficient base excision repair and if this occurs in a cell which already has defective HR, as a result of a BRCA mutation, then the combination of the two effects is lethal to the cell (Yap et al., 2010). PARPi are thus effective in cancers with mutated BRCA1 and BRCA2 genes (Audeh et al., 2010; Ledermann et al., 2011). This is a process known as synthetic lethality: the drugs block one DNA repair pathway whilst the BRCA mutation has already silenced another pathway (Dedes et al., 2011). A cell can survive with one pathway blocked but when both are affected the cell dies as there are many genes involved in this process.

Drugs may also be useful in other ovarian cancer patients, essentially those with tumours which are HR defective by other mechanisms. Recent work by our group (Mukhopadhyay et al., 2010) has suggested that up to 50% of ovarian cancers could be deficient in HR.

A simple test that would identify tumours with defective HR would therefore identify a group of patients who would likely respond to a PARP inhibitor.

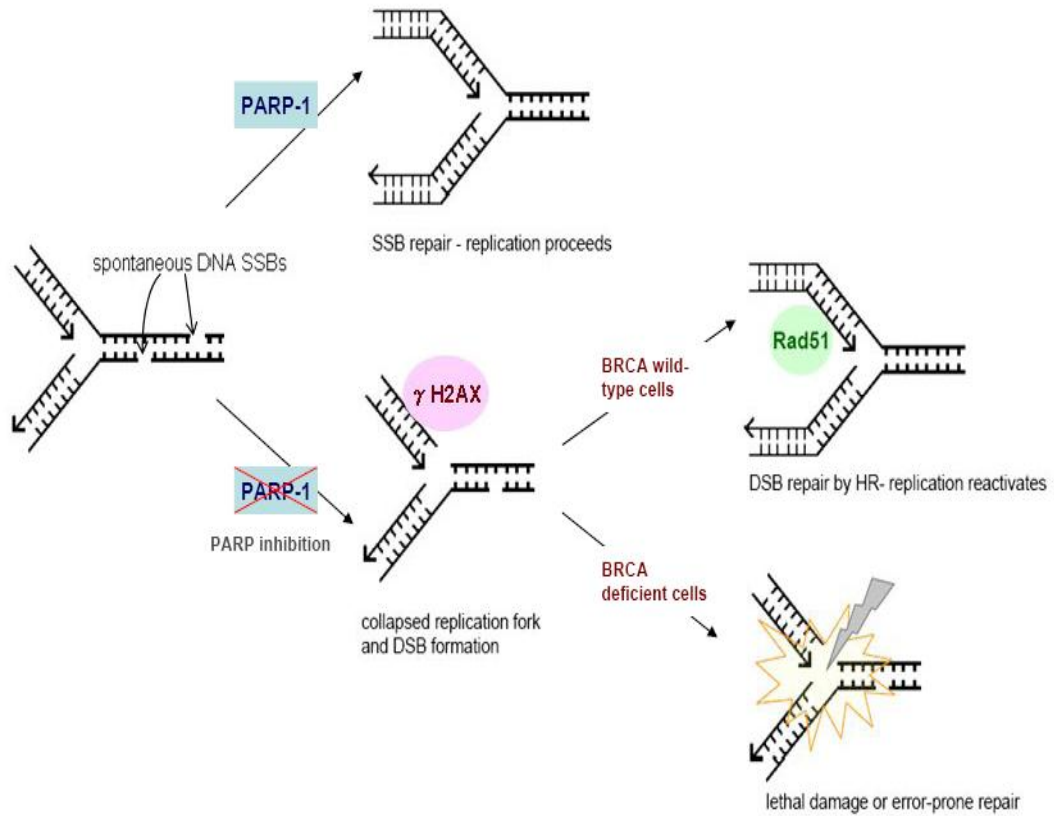


Figure 6-1 Selective cytotoxicity of PARP inhibitors

One pathway is silenced by using PARPi drug (top pathway). If a BRCA mutation or other genetic event silences the other pathway (bottom pathway) the cell dies.

6.3 Dataset

The dataset available for this analysis (Table 6-1) was developed by Gynaecological Oncology team at Queen Elizabeth hospital in Gateshead and NICR based in Newcastle University (Mukhopadhyay et al., 2010).

To collect the above dataset, ascities was collected at surgery from patients undergoing primary surgery for ovarian cancer. Primary cultures were generated to investigate the sensitivity to PARPi and collect DNA, mRNA and Protein. A functional assay was developed to measure HR (Mukhopadhyay et al., 2010) although this process works perfectly, it is very time consuming. Therefore it was decided that an investigation of the expression of relevant genes in this process may provide a simpler test, more suitable for clinical practice.

cDNA analysis became a widely used tool to study gene expression patterns in human cancer. In ovarian cancer this tool has been employed extensively to understand the different aspects of this cancer (Resnick et al., 2009; Gunawardana et al., 2009; Kulasingam et al., 2010).

For this study, eight primary cultures were selected for analysis. These had already been classified as HR competent (n=4) and HR deficient (n=4), using the functional assay described above. mRNA was extracted from unstimulated cells, converted into cDNA and analysed using the SA Biosciences DNA repair microarray chip which contains probes for 84 DNA repair genes. Data were normalised to housekeeping genes contained within the chip.

	AVG ΔC_t							
Gene	HR+	HR+	HR+	HR+	HR-	HR-	HR-	HR-
	PCO 44	PCO9 5	PCO9 6	PCO8 1	PCO9 0	PCO100	PC O93	PCO 106
APEX1	2.61	2.72	2.93	2.35	2.97	3.28	1.99	3.88
APEX2	8.1	8.85	8.43	8.33	9.53	8.74	8.18	10.05
ATM	8.8	8.45	8.55	8.69	9.18	7.62	7.35	8.63
ATR	6.98	6.83	6.81	7.38	7.42	7.25	6.03	8.02
ATXN3	7.3	8.38	7.77	7.29	8.28	7.24	6.89	8.05
BRCA1	7.86	6.46	8.44	8.77	7.74	8.92	11.9 5	8.53
BRCA2	9.67	6.64	9.12	9.67	9.59	12.76	9.9	9.43
BRIP1	6.17	6.68	7.38	7.33	6.27	6.98	6.85	7.99
CCNH	5.23	4.78	4.07	4.67	4.71	4.52	3.08	5.05
CCNO	10.1 7	10.74	11.13	10.84	10.84	11.34	8.54	11.64
CDK7	4.59	5.89	4.68	4.96	5.05	5.08	3.9	5.37
DDB1	3.31	3.4	2.84	2.82	3.95	2.4	2.16	4.29
DDB2	4.74	5.09	5.04	5.67	5.01	5.04	4.59	6.13
DMC1	11.0 5	11.87	11.18	12.36	11.68	14.48	13.6 3	12.91
ERCC1	5.86	5.76	5.09	5.55	7.94	5.78	5.07	6.52
ERCC2	5	5.09	4.1	5.12	5.6	3.93	4.09	5.46
ERCC3	6.52	6.52	6.09	6.47	6.84	5.96	5.35	7.59
ERCC4	7.33	7.26	6.8	6.29	7.32	6	5.33	8.38
ERCC5	4.08	4.72	3.96	4.49	4.04	4.04	3.35	4.84
ERCC6	8.19	8.27	7.84	8.25	8.36	7.09	6.31	8.62
ERCC8	7.78	7.72	7.04	7.25	7.49	6.62	6.56	7.52
EXO1	6.82	7.14	7.24	8.01	7.48	9.64	9.98	8.6
FEN1	4.97	5.07	4.73	5.51	6.47	6.36	5.23	7.3
LIG1	6.97	6.92	6.91	7.81	8.28	9.32	8.34	8.93
LIG3	6.65	7.33	6.28	6.26	7.69	5.74	5.87	7.95
LIG4	9.2	8.03	8.08	10.07	8.38	7.94	8.03	8.29
MGMT	4.19	4.71	3.55	3.81	4.5	4.05	3.59	4.09
MLH1	5.25	5.63	5.25	7.08	5.34	6.2	5.61	5.72

MLH3	7.23	7.45	6.94	7.28	7.72	6.9	5.79	6.83
MMS19	5.03	5.38	5.11	4.85	5.16	4.63	4.34	5.31
MPG	3.81	3.95	4.17	3.21	3.85	4.34	2.98	4.48
MRE11 A	10.2 4	9.92	9.41	10.23	10.26	10.43	9.23	11.3
MSH2	5.25	5.37	5.56	6.59	5.68	7.45	6.17	6.7
MSH3	6.54	6.95	6.26	6.37	6.94	6.74	5.86	7.39
MSH4	12.6 8	15.36	14.09	15.88	14.25	13.74	14.6 2	13.37
MSH5	8.46	9.29	8.25	8.8	10.05	8.9	7.84	9.74
MSH6	4.52	4.8	4.51	5.12	4.94	4.63	4.6	5.13
MUTYH	7.98	8.68	8.7	8.47	9.9	8.94	8.22	9.65
NEIL1	9	10.65	9.23	10.45	11.18	9.26	8.75	10.1
NEIL2	7.84	7.75	6.6	7.52	8.47	7.01	6.6	8.72
NEIL3	9.55	9.06	9.39	11.65	8.64	11.67	12.5 7	10.38
NTHL1	6.99	7.87	6.71	7.18	7.3	7.55	6.33	7.84
OGG1	7.16	7.06	6.75	6.83	7.51	6.59	6.29	7.7
PARP1	4.35	4.86	5.6	4.81	5.17	5.61	5.36	7.48
PARP2	5.47	6.45	6.04	6.36	6.16	7.13	5.59	6.66
PARP3	5.49	6.07	5.6	5.67	6.58	6	5.48	6.33
PMS1	6.5	7.27	6.74	7.1	7.27	6.8	6.16	7.48
PMS2	5.87	6.28	6.11	5.56	6.98	5.53	5.2	6.66
PNKP	6.21	6.37	5.55	6.13	7.56	5.73	4.99	7.55
POLB	5.96	6.09	5.85	6.4	5.88	6.28	5.15	6.35
POLD3	6.62	6.99	7.2	7.19	7.64	7.44	7.58	7.55
POLL	6.95	7.83	7.43	7.27	7.42	6.52	6.22	6.82
PRKDC	4.27	4.11	3.82	3.44	4.46	3.81	3.38	5.12
RAD18	6.11	6.74	5.63	6.33	6.11	6.52	5.28	7.14
RAD21	2.75	3.73	3.61	3.43	3.39	3.05	3.31	4.4
RAD23 A	4.33	4.87	4.26	4.07	4.98	3.7	3.44	4.77
RAD23 B	2.37	2.74	2.37	2.12	2.8	2.08	1.83	3.47
RAD50	6.99	7.61	7.09	8.03	8.28	7.19	6.84	8.32

RAD51	10.5 5	10.94	11.7	11.59	11.6	13.09	14.2 6	11.83
RAD51 C	6.56	8.06	6.98	7.78	7.25	8.06	7.09	8.48
RAD51 L1	8.05	8.32	7.94	8.03	8.67	7.99	7.1	8.82
RAD51 L3	9.13	8.44	8.29	9.43	9.37	8.78	9.36	9.52
RAD52	8.83	9.22	8.82	9.41	9.29	9.03	7.85	9.4
RAD54 L	7.37	7.99	8.34	9.69	8.2	10.06	9.26	8.63
RFC1	5.71	6.38	6.01	6.77	5.64	5.78	5.84	6.26
RPA1	4.04	3.85	3.75	3.77	4.24	4.06	3.35	4.88
RPA3	4.57	5.45	4.85	5.24	5.25	5.75	5.47	5.68
SLK	3.78	4.12	4.71	3.67	3.62	3.25	2.93	5.42
SMUG1	6.52	7.04	6.38	7.06	6.84	6.79	6.28	7.57
TDG	5.63	5.71	5.81	6.18	5.7	5.99	5.47	6.39
TOP3A	7.7	7.85	8.02	7.84	8.94	8.31	7.16	8.95
TOP3B	6.96	7.77	6.78	7.27	8.63	6.74	5.53	8.86
TREX1	6.15	6.67	7.18	7.08	7.72	7.1	6.81	6.88
UNG	4.82	5.85	5.07	5.55	5	5.36	4.81	5.73
XAB2	4.85	5.75	5.07	5.31	5.33	4.97	4.14	5.34
XPA	6.47	6.59	6.53	7.07	6.14	6.7	6.21	6.49
XPC	4.61	5.43	4.99	5.67	4.75	4.91	4.84	5
XRCC1	5.3	5.58	5.39	6	5.95	5.93	5.38	5.97
XRCC2	9.76	8.6	9.23	10.68	9.45	11.32	11.0 9	10.17
XRCC3	11.0 2	11.67	9.84	11.64	13.75	10.55	9.59	13.28
XRCC4	6.55	6.6	5.75	6.42	6.25	3.74	4	6.56
XRCC5	2.12	2.15	1.88	1.89	1.9	1.75	1.3	2.97
XRCC6	3.02	3.82	2.99	3.36	3.07	3.67	2.93	3.94
XRCC6 BP1	7.56	8.4	8.27	7.83	7.87	7.97	7.73	7.76

Table 6-1 Gene analysis dataset

6.4 Most important Genes

Previously it has been explained that the prediction of HR can play an important role in future treatment path of ovarian cancer. Using feature selection techniques (section 1.3.2), the analysis was performed to identify the most important genes for this prediction.

TPP (section 1.3.2.3) was used to analyse the dataset to compare the gene expression between HR competent and HR deficient tumours. This showed a clear separation between the two sets, Figure 6-2 illustrates the first view of the dataset produced using TPP. At first glance there is a separation between HR^- and HR^+ cases. HR^+ cases clustered together in the middle of the dataset and HR^- cases produced two separated clusters at the top and left of HR^+ cases. In order to discover the most important genes, TPP was further employed to generate the separated class view of the data.

Figure 6-3 illustrates the separated class view of the dataset using TPP. The HR^- and HR^+ cases are clearly separated from each other. HR^- cases clustered together at the top of the figure and HR^+ cases clustered together at the bottom of the figure.

In parallel with above analysis other feature selection techniques such as IG and GA (section 1.3.2) were employed for identification of the most important genes. The algorithms were used to identify the top 12 differentially regulated genes between the 4 HR competent and the 4 HR deficient tumours (Table 6-1). All three methods (TPP, GA and IG) identified LIG1 and POLD3 as the most important genes.

Discussion: Based on the TPP analysis there is a clear separation between HR^+ and HR^- cases. All three techniques identified LIG1 and POLD3 as the most important genes. The other genes that identified by all three techniques were FEN1 and EXO1. The results of TPP and GA have more similarities (10 of 12 important genes) compared to the results of IG.

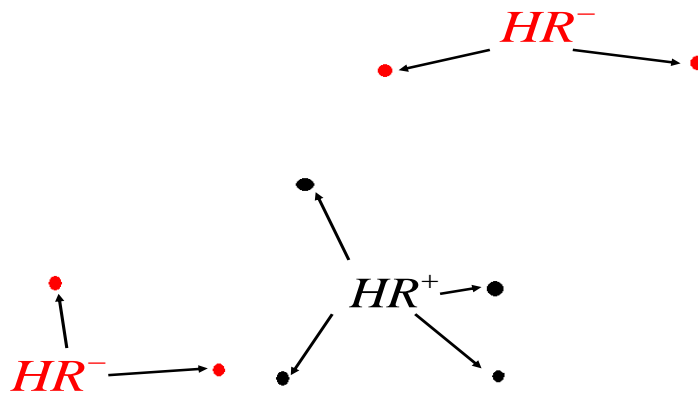


Figure 6-2 First view of dataset using TPP

The dots are displayed as a 2D projection of the 84 dimension analysis. HR^+ cases clustered in the middle whilst two HR^- cases clustered at the top and other two HR^- at the bottom left.

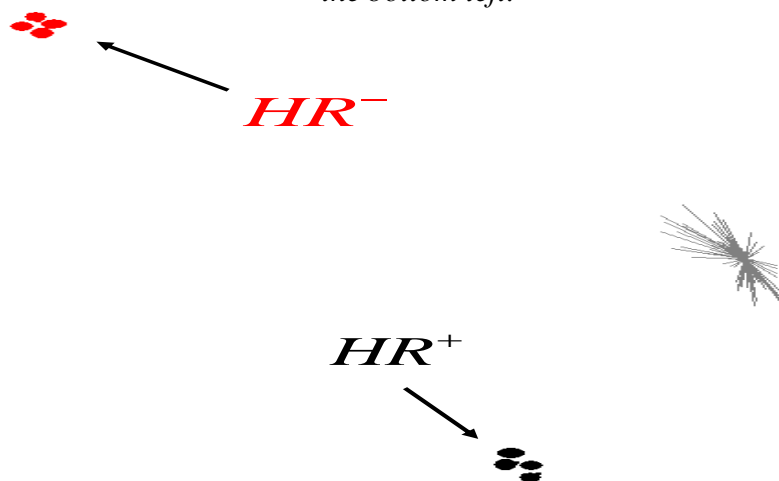


Figure 6-3 Separated class view of the dataset using TPP

This separated class's view of the dataset was produced using TPP and demonstrates that HR^- and HR^+ are clearly separated.

Method	Identified Genes
TPP	<u>LIG1, POLD3</u> , XPA, FEN1, EX01, XRCC4, LIG4, PARP3, RFC1, RPA3, POLL, RAD51
GA	<u>LIG1, POLD3</u> , XPA, EX01, PARP3, POLL, RFC1, FEN1, RPA3, RAD51, ERCC1, DMC1
IG	<u>LIG1, POLD3</u> , MLH3, MLH1, MMS19, MPG, FEN1, EX01, MGMT, LIG3, LIG4, MUTYH

Table 6-2 12 most differentially regulated genes identified between 4 HR competent and 4 HR deficient tumours using three methods

6.5 HR prediction models

The models used in previous sections (ANN, DT and BN) were then employed to develop prediction models of HR.

Table 6-2 summarizes the ANN results of prediction performance of HR. The accuracy of this model was 75%. The MAE value was 0.27 and RMSE value was 0.42. ANN managed to produce a very good model (AUC: 0.8750).

Figure 6-4 illustrates the ROC curves produces by ANN for prediction of HR. The AUC for both of the curves was 0.8750.

		Measurement	Result
		Accuracy	75.00%
		Mean absolute error	0.27
		Root mean squared error	0.42
Outcome	Recall	Precision	AUC
HR+	1.00	0.667	0.8750
HR-	0.750	1.00	0.8750

Table 6-3 ANN prediction performance results (HR)

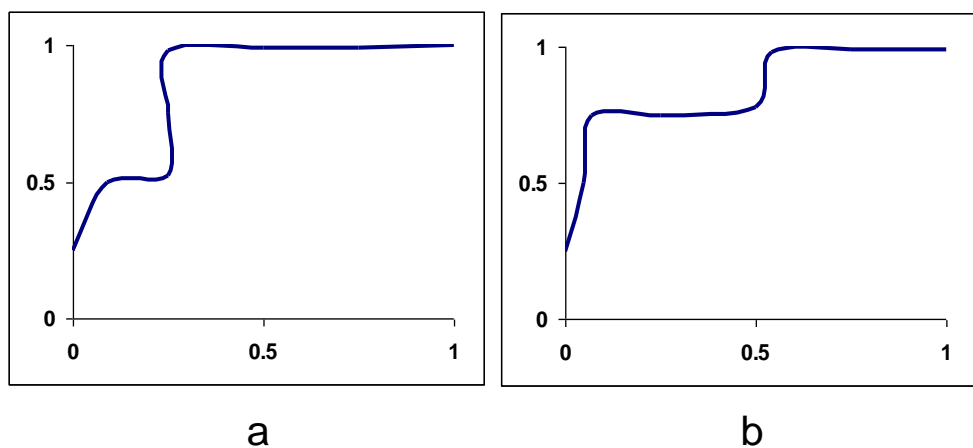


Figure 6-4 HR prediction ROC curves using ANN
a: HR+ and b: HR-

Table 6-3 summarizes the BN results of prediction performance of HR. The accuracy of this model was 62.50%. The MAE value was 0.46 and RMSE value 0.63. BN did not manage to produce any worthy model (AUC: 0.3750).

Figure 6-5 illustrates the ROC curves produced by BN for the prediction of HR. BN failed to produce any worthy model (AUC: 0.3750)

Measurement		Result	
Accuracy		62.50%	
Mean absolute error		0.46	
Root mean squared error		0.63	

Outcome	Recall	Precision	AUC
HR+	0.750	0.600	0.3750
HR-	0.500	0.633	0.3750

Table 6-4 BN prediction performance results (HR)

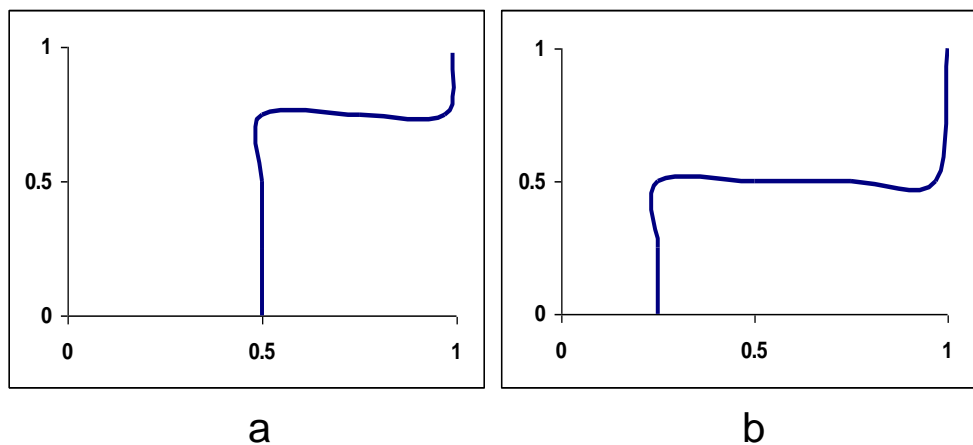


Figure 6-5 HR prediction ROC curves using BN
a: HR+ and b: HR-

Table 6-4 summarizes the DT results of prediction performance of HR. The accuracy of this model was 62.50%. The MAE value was 0.37 and RMSE value 0.61. DT managed to produce an average model (AUC: 0.6250).

Figure 6-6 illustrates the ROC curves produced by DT for the prediction of HR. DT managed to produce an average model (AUC: 0.6250).

Measurement		Result
Accuracy		62.50%
Mean absolute error		0.37
Root mean squared error		0.61

Outcome	Recall	Precision	AUC
HR+	0.750	0.600	0.6250
HR-	0.625	0.633	0.6250

Table 6-5 DT prediction performance results (HR)

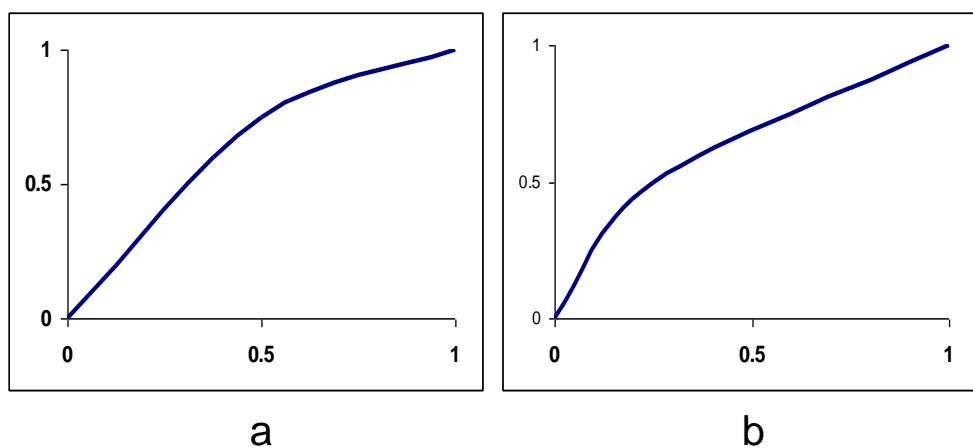


Figure 6-6 HR prediction ROC curves using DT
a: HR+ and b: HR-

Discussion: Based on the results ANN managed to develop a very good model (AUC: 0.8750). The accuracy of this model was 75%. The MAE value was 0.27 and RMSE value was 0.42. ANN managed to produce a very good model. The other models failed to introduce any good results. However DT performed better than BN.

6.6 Summary of the chapter

DNA is damaged many times during each cell cycle and these damages are repaired using DNA repair pathways. One of these repair pathways is HR and the two of the most important genes in these pathways are BRCA1 and BRCA2. The other DNA repair pathway is called base excision and PARP is an important protein in this pathway. A combination of defective HR and inhibited PARP is lethal to the cell. Therefore PARPi are effective in cancers with mutated BRCA1 and BRCA2 genes. As a result, a simple test that would identify tumours with defective HR would therefore identify a group of patients who would likely respond to a PARP inhibitor.

For this analysis a dataset which was developed by Gynaecological Oncology team at Queen Elizabeth Hospital in Gateshead and NICR based in Newcastle University is used. Briefly this dataset consist of eight primary cultures that had already been classified as HR competent (n=4) and HR deficient (n=4).

The feature selection techniques: TPP, IG and GA were employed for identification of the most important genes. The algorithms were used to identify the top 12 differently regulated genes between 84 available genes. The most important genes identified as LIG1 and POLD3.

Furthermore, ANN, BN and DT models were employed to predict the HR based on the available dataset that described previously. ANN outperformed other methods (Accuracy: 75%, AUC: 0.8750).

This analysis was performed using a small dataset and more cases are needed for analysis. Identified genes (LIG1 and POLD3) have to be prospectively validated using other conventional methods.

On the other hand, this analysis highlights that the test done on RNA is easier to roll out into clinical practice than the current gold standard HR test.

Chapter 7 Conclusion and Future Works

7.1 Introduction

This chapter draws conclusions on the work reported in this thesis. A brief summary of major contributions of this research is provided and suggestions for possible further studies are made.

7.2 Conclusions of the Thesis

This thesis presents an effective approach via machine learning and artificial intelligence to predict survival outcome in patients with ovarian cancer. The approach taken is novel in the field of ovarian cancer research. The main conclusions drawn from this thesis to answer the research questions are as follows:

- AI and ML models can effectively predict overall survival rates of ovarian cancer patients.
- Using AI and ML models made it possible to predict the outcome of the surgery.
- Feature selection techniques can identify the most important factors to predict survival rates and outcome of the surgery.

Some important conclusions derived from this study are as follows:

AI and ML models in most cases outperform conventional statistical models. ANN demonstrated the optimal performance of all the investigated models. The prediction performance of the models may improve by increasing the available data for analysis.

7.3 Summary of main contributions

The main goal of this research was to investigate and employ AI and ML models in order to provide more effective tools for clinicians managing patients with ovarian cancer. This goal leads to two major contributions, which are reviewed in related literature and the solution for the prediction.

7.3.1 The review of related problems

The review of the literature was divided into two sections: computing and medicine literature. The existing ovarian cancer literature shows that the dominant types of prediction models are statistical models. The AI and ML models were used to predict different aspects in other cancers; however the dominant prediction models in the field of ovarian cancer have all been statistical models until now.

The review of the computing literature indicates that three types of models are used for prediction. The ANN models as AI models, the BN as probabilistic models and DT as ML models are widely used.

The existing literature identifies several problems as follows:

- There is no effective way available for clinicians to measure the survival rates of the patients in an accurate way.
- Prediction of the outcome of the surgery remains an unsolved question and there is an essential need for this prediction.
- There is a need for an expert computer system to analyse the new biomarkers and treatment pathways.
- The developed models have to be easy to use.
- As the number of patient data increases the model has to cope with the new data and improve itself.

7.3.2 The solution for prediction

Developing a solution for predicting the survival rates and outcome of the surgery are the most important contribution of this research, as the prediction of these outcomes may have a significant effect on the quality of life of the patients. Accurate prediction of survival would allow patients to make informed treatment option decisions, faced with a poor prognosis patients may be more prone to opt for palliation, whereas patients with a good overall prognosis would be more likely to tolerate invasive and toxic treatment knowing that this will result in a survival benefit. Such accurate tools have not been available to date, although prognostic tools are available which can be applied to populations they are not accurate enough to allow use in individual cases. In this research the computation power of AI and ML models were demonstrated. Four main steps were used for the prediction: (a) collection of the relevant data; (b) development of variety of AI, ML and LR models; (c) analysis of the models by increasing the number of available data; (d) Identification of the most important variable for each prediction using all permutations search and feature selection methods;

The main capabilities of developed models for survival prediction are as follows:

The best developed model can predict the five year survival of the patient by 92.07% accuracy. The AUC of this model was 0.74. The accuracy of 92.07% is a very high

accuracy, however the AUC value is categorized as a good model and the model has to be improved to become an accurate enough model.

The model can predict the survival rate for 64 categories (section 4.3.1). Although the prediction accuracy for 64 categories is very low (based on the available data), by improving number of cases, this prediction may provide clinicians useful information. The prediction accuracy of the model increased as the number of cases improved.

The developed model outperforms the conventional statistical models. The prediction results of developed model are higher than currently available models in the field of ovarian cancer. The currently available model can only predict the five year survival however the developed model in this research can predict the survival for many categories.

Prediction is still not accurate enough for use on an individual patient but the encouraging thing is that the systems improve with data entry so this may develop into an accurate enough model eventually.

The main capabilities of developed models for outcome of the surgery prediction are as follows:

- The model can predict the outcome of the surgery for three main categories: (a) complete or optimal or sub-optimal cytoreduction; (b) complete or optimal cytoreduction versus sub-optimal cytoreduction; (c) complete cytoreduction versus optimal or sub-optimal cytoreduction.
- The developed model is a unique model for such prediction in the field of cancer research.

7.3.3 The prototype implementation and evaluation

The prototype of the models has been implemented using Java programming language in Windows environment (however it can be used on other platforms or web based). This prototype validates the approach used in this research and can be employed by clinicians for further use. The prototype provided a fundamental basis for conducting the experimental study of this research.

The developed prototype serves the following purposes:

- The easy to use graphical interface which is useable for any researcher.

- The capability of showing the survival rate and the prediction of the outcome of the surgery in discussed categories.
- The capability of adding and analysing new data to the dataset.
- The ability of the prototype to record newly added data.

7.4 Further work

There are some areas to explore in further details as follows:

- **Analysis of other feature sets**
 The datasets that were used in this research were taken from sequential patients who underwent primary surgery between 1995 and 2005 at the Northern Gynaecological Oncology Centre, Gateshead, UK, with appropriate ethical approval. For further work the suggestion is to expand the date range of the data that collected from this centre to include the data from 2005-2010. The gene analysis chapter (chapter 6) of thesis was based on a dataset that included 8 patients. The complexity of the micro array datasets requires the collection of new such data for further analysis. Furthermore, for complete validation of this research the datasets that collected from other centres around the world and UK has to be included. The reason is that other centres have different surgical approaches and so data from one centre may not reflect findings in another (would future models have to cope with this, the treating centre may be a very important feature in its own right) this will need exploring and validating in future work.
- **Including the missing values**
 As discussed previously (chapter 2) the datasets included a number of missing values. The original analysis was performed and the missing values were excluded from the datasets. For further research, it is suggested that such data be handled and included in the dataset. Semi supervised methods can be employed for handling the missing values.
- **Connection with DNA microarray dataset**
 The analysis of DNA micro array data discovered potential genes involved in abrogation of an important DNA repair pathway in ovarian cancer. It can be suggested for the next step of this research to explore the connection(s) between discovered genes and the OS and outcome of the surgery of the patients.
- **Exploration of other models**

This research included different models of ML and AI. There is a potential to expand these models to include further developed models in this (such as bootstrapping). The number of available new or improved models and packages (such as R) are growing, the investigation of new models are suggested.

- **Completion of the prototype**

The prototype may help clinicians to have a more clear idea of which treatment pathway to choose for a particular patient. Therefore, the completion of the prototype will be the next plan. This prototype (as discussed previously) may improve the quality of the life of the patients and a completed prototype may be employed by hospitals and clinicians for better management of ovarian cancer. It will also result in the generation of larger datasets which could be used for validating and further developing the models. Although the prototype has been made as easy to use as possible there will be a real challenge in encouraging busy clinicians to enter data in real time. As the validity of the prototype is demonstrated and the quality of the output data improve with time the hope is that the prototype becomes a clinically useful tool which clinicians will want to use. The hope is that the prototype act similar to ‘adjuvant online’, which is used all the time by breast cancer oncologists which doesn’t have the capability to capture any new data. Furthermore, the issues regarding the instituting this prototype in many centres such as security of data transfer and ethical considerations has to be investigated in future work.

- **Investigation of other survival data**

Cancer research scientists are investigating Event-free survival (EFS) and Relapse-free survival (RFS) for cancer patients. These factors were not included in the present analysis as the data were not available for this analysis and also the priority that was set for this research was to investigate the overall survival. It would be very beneficial to generate, analyse and include RFS and EFS into this analysis.

7.5 The overall achievement of the thesis

The overall achievements of this research can be summarized as follows:

The main achievement of this research is the presentation of a solution to predict the outcome in ovarian cancer. There are different models compared in this research and the most successful models were identified. Development of an AI system to predict the outcome of the surgery which has never been conducted before in the field of ovarian

cancer. Identification of the most important genes using TPP as a multivariate feature selection technique for the first time. Other achievement which is arguably novel, that is, developed model is able to predict the survival rates of the patients for 64 categories based on the median maximum survival value in the dataset. By improving the number of cases available in the dataset this model can effectively predict the survival rates in the range of a month. Lastly, this research presents a new pathway of using AI systems for helping clinicians to explore different treatment options for a patient.

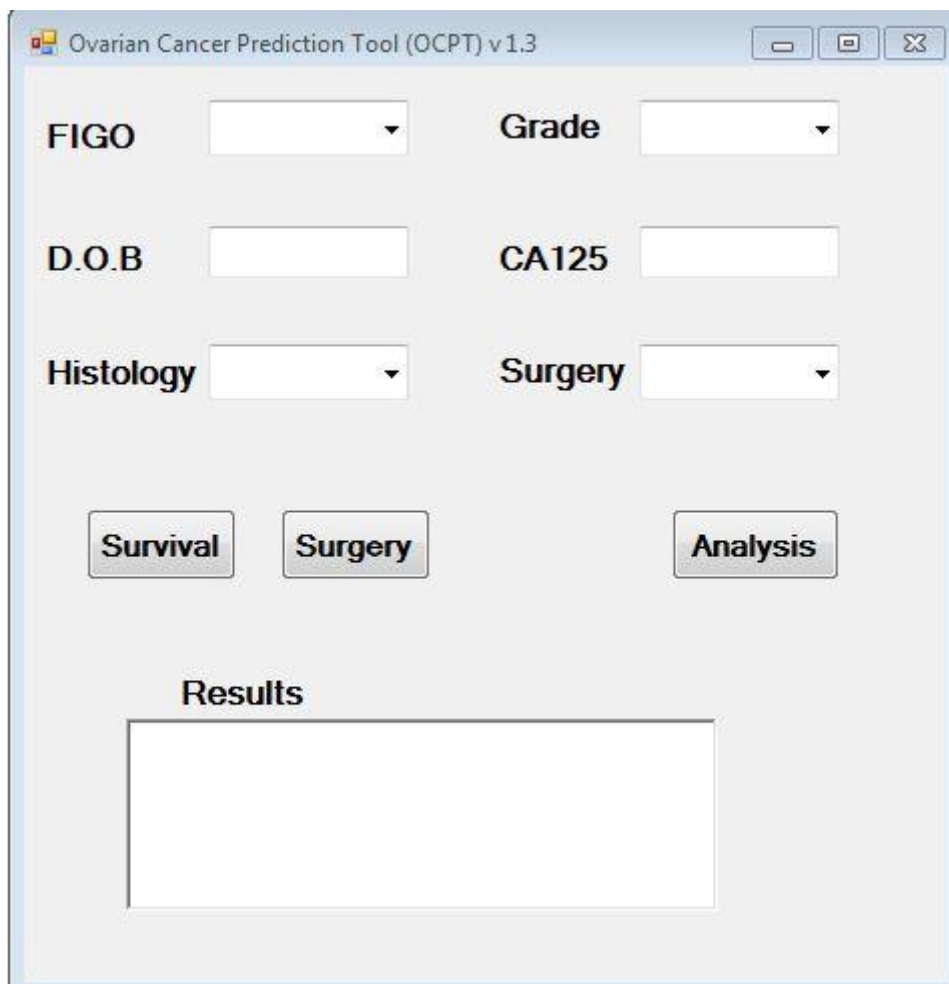
Overall, present research makes advances in the clinical decisions making area in the medicine. The developed models make a strong argument that such models can be employed by health care teams to explore different treatments for a patient. Furthermore, this approach has also demonstrated its efficiency in dealing with large amounts of clinical data and finding the associations between the different markers to predict outcomes in ovarian cancer. We believe that the findings presented in this thesis will draw more attention to the area and attract more research in this field.

Appendices

Appendix 1 Prototype snapshot

Following is the snap shot of the main page of the developed prototype in this research.

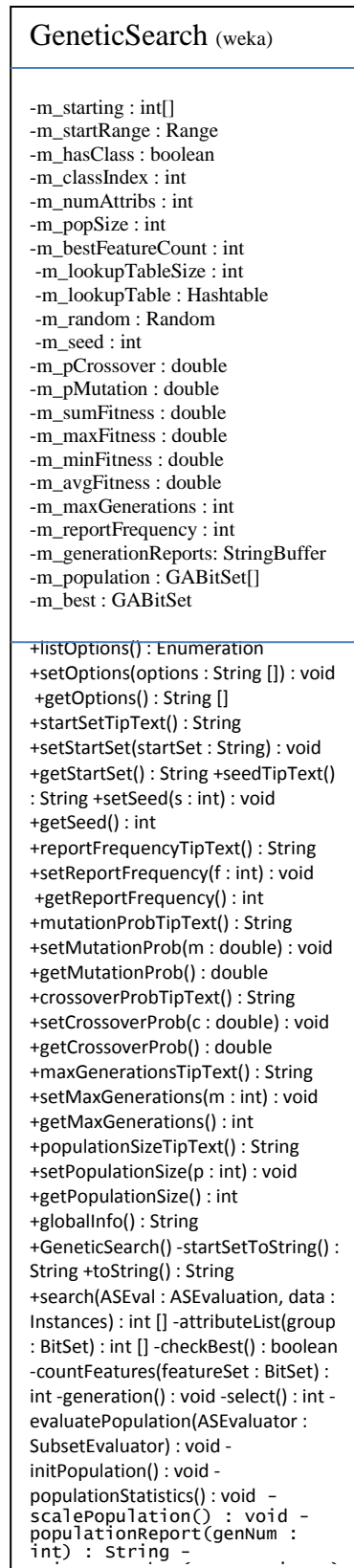
- 1- User enters the data into the prototype
- 2- User selects survival/surgery for the prediction of the survival/surgery.
- 3- After the analysis user has the option to include this new data into the dataset.
- 4- By using analysis option user can produce different results (ROC curves, LR prediction, other modelling results, ...)



The screenshot shows a software window titled "Ovarian Cancer Prediction Tool (OCPT) v1.3". The interface includes several input fields and buttons. On the left side, there are dropdown menus for "FIGO", "Histology", and "Grade", and text input fields for "D.O.B" and "CA125". On the right side, there are dropdown menus for "Surgery" and "Grade", and a text input field for "CA125". Below the input fields, there are three buttons: "Survival", "Surgery", and "Analysis". At the bottom of the window, there is a section labeled "Results" with a large empty rectangular box for displaying the output.

Appendix 2 class diagrams

Following are some of the class diagrams for the (weka) codes that used in this research.



InfoGain (weka)

-m_missing_merge : Boolean
-m_Binarize : boolean
-m_InfoGains : double[]

+globalInfo() : String
+InfoGainAttributeEval()
+listOptions() : Enumeration
+setOptions(options : String []) : void
+getOptions() : String []
+binarizeNumericAttributesTipText() : String
+setBinarizeNumericAttributes(b : boolean) : void
+getBinarizeNumericAttributes() : boolean
+missingMergeTipText() : String
+setMissingMerge(b : boolean) : void
+getMissingMerge() : boolean
+buildEvaluator(data : Instances) : void
#resetOptions() : void
+evaluateAttribute(attribute : int) : double
+toString() : String

NaiveBayes(weka)

#m_Distributions : Estimator[][]
#m_ClassDistribution : Estimator
#m_UseKernelEstimator : boolean = false
#m_UseDiscretization : boolean = false
#m_NumClasses : int
#m_Instances : Instances
#DEFAULT_NUM_PRECISION : double = 0.01
#m_Disc : Discretize = null

+buildClassifier(instances : Instances) : void
+updateClassifier(instance : Instance) : void
+distributionForInstance(instance : Instance) : double []
+listOptions() : Enumeration
+setOptions(options : String []) : void
+getOptions() : String []
+toString() : String
+getUseKernelEstimator() : boolean
+setUseKernelEstimator(v : boolean) : void
+getUseSupervisedDiscretization() : boolean
+setUseSupervisedDiscretization(new blah : boolean) : void

PCA (weka)

m_trainInstances : Instances
-m_trainCopy : Instances
-m_transformedFormat : Instances
-m_originalSpaceFormat : Instances
-m_hasClass : boolean
-m_classIndex : int
-m_numAttribs : int
-m_numInstances : int
-m_correlation : double[][]
-m_eigenvectors : double[][]
-m_eigenvalues : double[] = null
-m_sortedEigens : int[]
-m_sumOfEigenValues : double = 0.0
-m_replaceMissingFilter :
ReplaceMissingValues
-m_normalizeFilter : Normalize
-m_nominalToBinFilter :
NominalToBinary
-m_attributeFilter : Remove
-m_attrFilter : Remove
-m_outputNumAtts : int = -1
-m_normalize : boolean = true
-m_coverVariance : double = 0.95
-m_transBackToOriginal : boolean =
false
-m_eTranspose : double[][]

+globalInfo() : String
+listOptions() : Enumeration
+setOptions(options : String []) : void
-resetOptions() : void
+normalizeTipText() : String
+setNormalize(n : boolean) : void
+getNormalize() : boolean
+varianceCoveredTipText() : String
+setVarianceCovered(vc : double) :
void
+getVarianceCovered() : double
+transformBackToOriginalTipText()
: String
+setTransformBackToOriginal(b :
boolean) : void
+getTransformBackToOriginal() :
boolean
+getOptions() : String []
+buildEvaluator(data : Instances) :
void
-buildAttributeConstructor(data :
Instances) : void
+transformedHeader() : Instances
+transformedData() : Instances
+evaluateAttribute(att : int) : double
-fillCorrelation() : void
-principalComponentsSummary() :
String
+toString() : String
-matrixToString(matrix : double [][])
: String
-convertInstanceToOriginal(inst :
Instance) : Instance
+convertInstance(instance : Instance)
: Instance
-setOutputFormatOriginal() :
Instances
-setOutputFormat() : Instances

ANN (weka)
<pre> -m_instances : Instances -m_currentInstance : Instan -m_numeric : boolean -m_attributeRanges : doub -m_attributeBases : double -m_numClasses : int = 0 -m_numAttributes : int = 0 -m_nextId : int -m_selected : FastVector -m_graphers : FastVector -m_numEpochs : int -m_stopIt : boolean -m_stopped : boolean -m_accepted : boolean -m_win : JFrame -m_autoBuild : boolean -m_gui : boolean -m_valSize : int -m_driftThreshold : int -m_randomSeed : long -m_random : Random -m_nominalToBinaryFilter : -m_hiddenLayers : String -m_normalizeAttributes : bo -m_decay : boolean -m_learningRate : double -m_momentum : double -m_epoch : int -m_error : double -m_reset : boolean -m_normalizeClass : boole -m_outputs : NeuralEnd[] -m_inputs : NeuralEnd[] -m_neuralNodes : NeuralC -m_nodePanel : NodePane -m_controlPanel : ControlP -m_sigmoidUnit : SigmoidU -m_linearUnit : LinearUnit +NeuralNetwork() +setDecay(d : boolean) : void +getDecay() : boolean +setReset(r : boolean) : void +getReset() : boolean +setNormalizeNumericClass(c : boolean) : void +getNormalizeNumericClass() : boolean +setNormalizeAttributes(a : boolean) : void +getNormalizeAttributes() : boolean +setNominalToBinaryFilter(f : boolean) : void +getNominalToBinaryFilter() : boolean +setRandomSeed(l : long) : void +getRandomSeed() : long +setValidationThreshold(t : int) : void +getValidationThreshold() : int +setLearningRate(l : double) : void +getLearningRate() : double +setMomentum(m : double) : void +getMomentum() : double +setAutoBuild(a : boolean) : void +getAutoBuild() : boolean +setHiddenLayers(h : String) : void +getHiddenLayers() : String +setGUI(a : boolean) : void +getGUI() : boolean +setValidationSetSize(a : int) : void +getValidationSetSize() : int +setTrainingTime(n : int) : void +getTrainingTime() : int -addNode(n : NeuralConnection) : </pre>

References

- Abbod, M. F., Catto, J. W. F., Linkens, D. A. and Hamdy, F. C. (2007) 'Application of artificial intelligence to the management of urological cancer', *The Journal of Urology*, 178(4), pp. 1150-1156.
- Adali, T. and Haykin, S. (2010) *Adaptive Signal Processing: Next Generation Solutions*. New Jersey: Wiley-Blackwell.
- Adam, B. L., Qu, Y. and et al. (2002) 'Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men', *Cancer Res*, 62, pp. 3609-3614.
- AIHW (2006) *Ovarian cancer in Australia: an overview*. Available at: <http://aihw.gov.au/publications/can/oca06/oca06.pdf> (Accessed: 26-11-2008).
- Al Shalabi, L., Najjar, M. and Al Kayed, A. (2006) 'A framework to deal with missing data in data sets', *Journal of Computer Science*, 2(9), pp. 740-745.
- Allen, D. G. (2010) 'The management of epithelial ovarian cancer: Neoadjuvant chemotherapy and interval surgery', *South Afr J Gynaecol Oncol*, 2(2), pp. 67-68.
- Alpaydin, E. (2004) *Introduction to Machine Learning (Adaptive Computation and Machine Learning)*. Cambridge: MIT Press.
- Andrew, A. S., Nelson, H. H., Kelsey, K. T. and et al. (2006) 'Concordance of multiple analytical approaches demonstrates a complex relationship between DNA repair gene SNPs, smoking and bladder cancer susceptibility', *Carcinogenesis* 27, pp. 1030-7.
- Ashraf, M., Le, K. and Huang, X. (2010) 'Information gain and adaptive neuro-fuzzy inference system for breast cancer diagnoses', *Computer Sciences and Convergence Information Technology (5th International Conference on)*, pp. 911-915.
- Asyali, M. H., Colak, D., Demirkaya, O. and Inan, M. S. (2006) 'Gene expression profile classification: A review', *Current Bioinformatics*, , 1, pp. 55-73.
- Audeh, M. and et al. (2010) 'Oral poly(ADP-ribose) polymerase inhibitor olaparib in patients with BRCA1 or BRCA2 mutations and recurrent ovarian cancer: a proof-of-concept trial', *Lancet*, 376, pp. 245-251.
- Ayer, T., Alagoz, O. and et al. (2010) 'Breast cancer risk estimation with artificial neural networks revisited: Discrimination and calibration', *Cancer*, 116(14), pp. 3310-3321.
- Baldi, P. and Brunak, S. (2001) *Bioinformatics: The Machine Learning Approach*. 2nd edn. Cambridge: MIT Press.

- Balega, J. and Shepherd, J. H. (2007) 'Surgical Management of Patients with Epithelial Ovarian Cancer', in Reznick, R. (ed.) *Cancer of the ovary*. Cambridge: Cambridge University Press.
- Banks, D. L., House, L. L., McMorris, F., Arabie, P. and Gaul, W. (2004) *Classification, Clustering, and Data Mining Applications*. New York: Springer-Verlag Berlin and Heidelberg GmbH & Co.
- Banning, M. (2008) 'A review of clinical decision making: models and current research', *J Clin Nurs*, 17(2), pp. 187-195.
- Barnett, J. C., Bean, S. M., Whitaker, R. S., Kondoh, E., Baba, T., Fujii, S. and et al. (2010) 'Ovarian cancer tumour infiltrating T-regulatory (T(reg)) cells are associated with a metastatic phenotype', *Gynecol Oncol* 116, pp. 556-562.
- Barnett, V. and Lewis, T. (1994) *Outliers in Statistical Data*. 3rd edn. London: John Wiley & Sons.
- Barni, M., Buti, F., Bartolini, F. and Cappellini, V. (2000) 'A quasi-Euclidean norm to speed up vector median filtering', *IEEE Transactions on Image Processing*, 9(10), pp. 1704-1709.
- Beer, D. G., Kardias, S. L., Huang, C. C., Giordano, T. J. and et al. (2002) 'Gene-expression profiles predict survival of patients with lung adenocarcinoma', *Nature Medicine*, 8, pp. 816-824.
- Ben-Dor, A., Bruhn, L., Friedman, N., Nachman, I., Schummer, M. and Yakhini, Z. (2000) 'Tissue classification with gene expression profiles', *Journal of Computational Biology*, 7(2), pp. 559-584.
- Benner, P. (1982) 'From novice to expert', *Amer.J.Nursing*, 82(1), pp. 402-407.
- Benner, P. (1984) *From Novice to Expert: Excellence and Power in Clinical Nursing Practice*. Massachusetts: Addison-Wesley.
- Berrar, D., Sturgeon, B., Bradbury, I. and Dubitzky, W. (2003) 'Microarray data integration and machine learning techniques for lung cancer survival prediction', *Proceedings of the CAMDA-2003*.
- Bicciato, S., Luchini, A. and Di Bello, C. (2002) 'PCA disjoint models for multiclass cancer analysis using gene expression data', *Bioinformatics* 19(5), pp. 571-578.
- Bioinformatics (2008) *Gene Expression Analysis*. Available at: http://bioinformatics.org/BI201A_Gene_Expression_Analysis.
- Bishop, C. M. (1996) *Neural Networks for Pattern Recognition*. New York: Oxford University Press.

- Blazeby, J. M., Wilson, L., Metcalfe, C., Nicklin, J., English, R. and Donovan, J. L. (2006) 'Analysis of clinical decision-making in multi-disciplinary cancer teams', *Ann Oncol*, 17, pp. 457-460.
- Blum, A. and Langley, P. (1997) 'Selection of relevant features and examples in machine learning', *Artificial Intelligence*, 97(1-2), pp. 245-271.
- Bolstad, W. (2004) *Introduction to Bayesian Statistics*. New York: Wiley.
- Bourne, T. H., Campbell, S., Reynolds, K. and et al. (1994) 'The potential role of serum CA125 in an ultrasound-based screening program for familial ovarian cancer', *Gynecol. Oncol.*, 52, pp. 379-385.
- Boutros, P. C., Lau, S. K., Pintilie, M. and al., e. (2009) ' Prognostic gene signatures for non-small-cell lung cancer', *Proc Natl Acad Sci USA*, 106.
- Braga-Neto, U., Hashimoto, R., Dougherty, E. R., Nguyen, D. V. and Carroll, R. J. (2004) 'Is cross-validation better than resubstitution for ranking genes?', *Bioinformatics*, 20, pp. 253-265.
- Brazma, A. and Vilo, J. (2000) 'Gene expression data analysis', *FEBS Letters*, 480, pp. 17-24.
- Bristow, R. E. and Chi, D. S. (2006) 'Platinum-based neoadjuvant chemotherapy and interval surgical cytoreduction for advanced ovarian cancer: a meta-analysis', *Gynecol Oncol*, 103, pp. 1070-1076.
- Bristow, R. E., Duska, L. R., Lambrou, N. C. and et al. (2000) 'A model for predicting surgical outcome in patients with advanced ovarian carcinoma using computed tomography', *Cancer*, 89, pp. 1532-1540.
- Bristow, R. E. and et al. (2002) 'Survival Effect of Maximal Cytoreductive Surgery for Advanced Ovarian Carcinoma During the Platinum Era: A Meta-Analysis', *J Clin Oncol*, 20(5), pp. 1248-1259.
- Bulashevskaya, S., Szakacs, O., Brors, B., Eils, R. and Kovacs, G. (2004) 'Pathways of urothelial cancer progression suggested by Bayesian network analysis of allelotyping data', *Int J Cancer*, 110, pp. 850-856.
- Burke, H. B., Goodman, P. H., Rosen, D. B., Henson, D. E. and Weinstein, J. N. (2001) ' Artificial neural networks improve the accuracy of cancer survival prediction', *Cancer*, 91, pp. 857-862.
- Burnside, E., Rubin, D. and Shachter, R. (2004) 'Using a Bayesian network to predict the probability and type of breast cancer represented by microcalcifications on mammography', *Proceedings of the 11th World Congress on Medical Informatics* San Francisco.
- Butz, M. V., Goldberg, D. E. and Lanzi, P. L. (2005) 'Gradient descent methods in learning classifier systems: improving XCS performance in multistep problems', *IEEE Transactions on Evolutionary Computation*, 9(5), pp. 452-473.

- Cabestany, J. and Prieto, A. (2006) *Artificial Neural Networks*. Berlin: Springer.
- Caipeng, W., Deng, J. and Yang, Y. (2010) 'A Research of Fuzzy Neural Network in Ferromagnetic Target Recognition', in Zeng, Z. and Wang, J. (eds.) *Advances in Neural Network Research and Applications*. Berlin: Springer, pp. 148-156.
- Cancer Research UK (2007) *Artificial intelligence could cut hours from radiation treatment*. Available at: <http://info.cancerresearchuk.org/news/archive/newsarchive/2007/february/18056637>.
- Cancer Research UK (2008b) *Statistics and outlook for ovarian cancer*. Available at: <http://www.cancerhelp.org.uk/help/default.asp?page=5449> (Accessed: 29-11-2008).
- Cancer Research UK (2008a) *UK Ovarian Cancer incidence statistics*. Available at: <http://info.cancerresearchuk.org/cancerstats/types/ovary/incidence/#source5> (Accessed: 29-11-2008).
- Cancer Research UK (2008c) *UK Ovarian Cancer incidence statistics*. Available at: <http://www.cancerhelp.org.uk/help/default.asp?page=143> (Accessed: 29-11-2008).
- Catto, J. W., Linkens, D. A., Abbod, M. F., Chen, M., Burton, J. L., Feeley, K. M. and Hamdy, F. C. (2003) 'Artificial intelligence in predicting bladder cancer outcome: a comparison of neuro-fuzzy modeling and artificial neural networks', *Clin Cancer Res*, 9, pp. 4172-4177.
- Cerbinskaite, A. and et al. (2011) 'Defective homologous recombination in human cancers', *Cancer Treatment Reviews*, 37.
- Chakraborty, A. and Maka, H. (2005) 'Biclustering of Gene Expression Data Using Genetic Algorithm', *Proceedings of the 2005 IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology*, pp. 1-8.
- Chang, E. Y., Hop, S. C. H., Wang, X., Max, W.-Y. and Lyu, M. R. (2005) 'A unified learning paradigm for large-scale personalized information management', *Emerging Information Technology Conference*, pp. 4-8.
- Changjing, S. and Shen, Q. (2005) 'Aiding Classification of Gene Expression Data with Feature Selection: A Comparative Study', *International Journal of Computational Intelligence Research*, 1(1), pp. 68-76.
- Chapelle, O., Vapnik, V., Bousquet, O. and Mukherjee, S. (2002) 'Choosing multiple parameters for support vector machines', *Machine Learning*, 4(1), pp. 131-159.

- Cheng, J. and Li, Q. S. (2008) 'Reliability analysis of structures using artificial neural network based genetic algorithms', *Comput. Methods Appl. Mech. Engrg*, 197(45-48), pp. 3742-3750.
- Chi, D. S. and Hoskins, W. J. (2000) 'Primary Surgical Management of Ovarian Cancer', in Bartlett, J. (ed.) *Ovarian Cancer Methods and Protocols*. New Jersey: Humana Press.
- Chi, D. S., Palayekar, M. J., Sonoda, Y., Abu-Rustum, N. R., Awtrey, C. S., Huh, J. and et al. (2008) 'Nomogram for survival after primary surgery for bulky stage IIIC ovarian carcinoma', *Gynecol Oncol*, 108, pp. 191-194.
- Cho, S. and Ryu, J. (2002) 'Classifying gene expression data of cancer using classifier ensemble with mutually exclusive features', *Proceedings of the IEEE*, 90(11), pp. 1744-1753.
- Chu, C. S. and Rubin, S. C. (2005) 'Epidemiology, staging and clinical characteristics', in Bristow, R. E. and Karlan, B. Y. (eds.) *Surgery for Ovarian Cancer: Principles and Practice*. Oxon: Taylor & Francis, pp. 72-83.
- Cianfranco, M. and Goldstein, L. J. (2004) 'Prognostic and predictive factors in early-stage breast cancer', *Oncologist* 9, pp. 606-616.
- Coppin, B. (2004) *Artificial Intelligence Illuminated*. Sudbury: Jones and Bartlett Publishers.
- Crijns, A. P., Fehrmann, R. S., De Jong, S., Gerbens, F., Meersma, G. J. and et al. (2009) 'Survival-related profile, pathways, and transcription factors in ovarian cancer', *PLoS Med*, 6.
- Daelemans, W., Bosch, A. and Zavrel, J. (1999) 'Forgetting Exceptions is Harmful in Language Learning', *Machine Learning*, 34(1).
- Dechter, R. and Mateescu, R. (2004) "Mixtures of deterministic-probabilistic networks and their AND/OR search space", *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 70, pp. 120-129.
- Dedes, K. J., Wilkerson, P. M., Wetterskog, D., Weigelt, B., Ashworth, A. and Reis-Filho, J. S. (2011) 'Synthetic lethality of PARP inhibition in cancers lacking BRCA1 and BRCA2 mutations', *Cell Cycle*, 10, pp. 1192-1199.
- Diaconis, P. and Freedman, D. (1984) 'Asymptotes of Graphical Projection Pursuit', *Annals of Statistics*, 12, pp. 793-815.
- Dietterich, T. G. (1998) 'Approximate statistical tests for comparing supervised classification learning algorithms', *Neural Comput.*, 10(7), pp. 1895-1923.
- Djavan, B., Remzi, M. and et al. (2002) 'Novel artificial neural network for early detection of prostate cancer', *Clin Oncol*, 20, pp. 921-929.

- Dobbin, K., Zhao, Y. and Simon, R. (2008) 'How large a training set is needed to develop a classifier for microarray data', *Clin Cancer Res*, 14, pp. 108-114.
- Dragonieri, S., Annema, J. T., Schot, R. and et al. (2009) 'An electronic nose in the discrimination of patients with non-small cell lung cancer and COPD', *Lung Cancer*, 64, pp. 166-170.
- Dubitzky, W., Granzow, M. and Berrar, D. P. (2006) *Fundamentals of Data Mining in Genomics and Proteomics*. New York: Springer-Verlag.
- Duda, R. O., Hart, P. E. and Stork, D. G. (2000) *Pattern Classification*. 2nd edn. New York: Wiley-Interscience.
- Dudoit, S., Fridlyand, J. and Speed, T. P. (2002) 'Comparison of discrimination methods for the classification of tumours using gene expression data', *Journal of the American Statistical Association*, 97(457), pp. 77-87.
- Dziuda, D. M. (2010) *Data Mining for Genomics and Proteomics: Analysis of Gene and Protein Expression Data*. New Jersey: Wiley-Blackwell.
- Edmondson, R. J. and Todd, A. R. (2008) 'Ovarian cancer', *International Encyclopedia of Public Health*, 4, pp. 712-718.
- Efron, B. (1983) 'Estimating the error rate of a prediction rule: improvements on crossvalidation', *J. Amer. Stat. Ass.*, 78, pp. 316-331.
- Elmasry, K. and Gayther, S. A. (2007) 'Epidemiology of Ovarian Cancer', in Reznick, R. (ed.) *Cancer of the ovary*. Cambridge: Cambridge University Press.
- Enas, G. G. and Choi, S. C. (1986) 'Choice of the smoothing parameter and efficiency of the k-nearest neighbour classification', *Comput. Math. Applic.*, 12, pp. 235-244.
- Engelbrecht, A. (2007) *Computational Intelligence: An Introduction*. 2nd edn. London: John Wiley and Sons.
- Enshaie, A. and Faith, J. (2009) *Data Exploration using Targeted Projection Pursuit*. Northumbria.
- Faith, J. (2007) 'Targeted Projection Pursuit for Interactive Exploration of High-Dimensional Data Sets', *Proceedings of the 11th International Conference Information Visualization*, pp. 286-292.
- Faith, J., Mintram, R. and Angelova, M. (2006) 'Targeted projection pursuit for visualizing gene expression data classifications', *Bioinformatics*, 22(21), pp. 2667-2673.
- Farley, J. and Birrer, M. J. (2010) 'Discovery of Novel Targets', in Kaye, S., Brown, R., Gabra, H. and Gore, M. (eds.) *Emerging Therapeutic Targets in Ovarian Cancer*. London: Springer

- Fausett, L. (1994) *Fundamentals of Neural Networks*. NJ: Prentice-Hall.
- Fedorov, V., Mannino, F. and Zhang, R. (2009) 'Consequences of dichotomization', *Pharm Stat*, 8, pp. 50-61.
- Fogel, D. B., Wasson III, E. C. and M., B. E. (1995) 'Evolving neural networks for detecting breast cancer', *Cancer Letters*, 96, pp. 49-53.
- Fonseca, C. and Fleming, P. (1995) 'An overview of evolutionary algorithms in multiobjective optimization', *Evolutionary Computation*, 3(1), pp. 1-16.
- Forrest, S. (1993) 'Genetic algorithms: principles of natural selection applied to computation', *Science*, 261, pp. 872-878.
- Friedman, C. (2009) 'Discovering Novel Adverse Drug Events Using Natural Language Processing and Mining of Electronic Health Record', in Combi, C., Shahar, Y. and Abu-Hanna, A. (eds.) *Artificial Intelligence in Medicine: 12th Conference on Artificial Intelligence in Medicine, AIME 2009, Verona, Italy*. Berlin: Springer.
- Friedman, J. H. and Tukey, J. W. (1974) 'A projection pursuit algorithm for exploratory data analysis', *IEEE Transactions on Computers*, 23, pp. 881-890.
- Fujikoshi, Y., Ulyanov, V. V. and Shimizu, R. (2010) *Multivariate Statistics: High Dimensional and Large-Sample Approximations*. New York: John Wiley & Sons.
- Funt, S. A., Hricak, H., Abu-Rustum, N., Mazumdar, M., Felderman, H. and Chi, D. S. (2004) 'Role of CT in the management of recurrent ovarian cancer', *AJR Am J Roentgenol*, 182(393-398).
- Gabrilovich, E. and Markovitch, S. (2004) 'Text categorization with many redundant features: Using aggressive feature selection to make SVMs competitive with C4.5', *In Proceedings of the Twenty-First International Conference on Machine learning*, pp. 321-328.
- Gaten, T. (2000) *Normal distributions*. Available at: <http://www.le.ac.uk/bl/gat/virtualfc/Stats/normal.htm>.
- Gedeon, T. D., Wong, P. M. and Harris, D. (1995) 'Balancing Bias and Variance: Network Topology and Pattern Set Reduction Techniques', in Mira, J. and Sandoval, F. (eds.) *Proceedings of the International Workshop on Artificial Neural Networks, Lecture Notes in Computer Science*. pp. 551-558.
- Gelman, A., Carlin, J. B., Stern, H. S. and Rubin, D. B. (1995) *Bayesian data analysis*. London: Chapman & Hall.

- Gerestein, C. G., van der Spek, D. W. and et al. (2009) 'Prediction of residual disease after primary cytoreductive surgery for advanced-stage ovarian cancer: accuracy of clinical judgment', *Int J Gynecol Cancer*, 19(9), pp. 1511-1515.
- Gesu, V., Giancarlo, R., Bosco, G. L., Raimondi, A. and Scaturro, D. (2005) 'GenClust: A Genetic Algorithm for Clustering Gene Expression Data', *BMC Bioinformatics*, 6(1), p. 289.
- Gevaert, O., De Smet, F., Timmerman, D., Moreau, Y. and De Moor, B. (2006) 'Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks', *Bioinformatics* 22(14), pp. 184-190.
- Ghaemmghami, S., Orton, K. and Soutter, P. (2003) *Key Advances in the Clinical Management of Ovarian Cancer*. London: Royal Society of Medicine Press.
- Goldberg, D. (1989) *Genetic Algorithms in Search, Optimization in Machine Learning*. Boston: Addison-Wesley Longman Publishing.
- Golub, J. and Loan, C. F. (1999) *Matrix Computations, Johns Hopkins Studies in the Mathematical Sciences*. Baltimore: Johns Hopkins University Press.
- Greer, B. T. and Khan, J. (2004) 'Diagnostic classification of cancer using DNA microarrays and artificial intelligence', *Ann N Y Acad Sci*, 1020, pp. 49-66.
- Gunawardana, C. G., Memari, N. and Diamandis, E. P. (2009) 'Identifying novel autoantibody signatures in ovarian cancer using high-density protein microarrays', *Clin. Biochem.*, 42, pp. 426-429.
- Haddow, C., Perry, J., Durrant, M. and Faith, J. (2008) 'Functional analysis of the primary structure of proteins using vector representations of amino acid property sequences', *IJCA*, 5(1), pp. 50-58.
- Han, J., Kamber, M. and Pei, J. (2006) *Data Mining: Concepts and Techniques*. 2nd edn. San Francisco: Morgan Kaufmann.
- Hartmann, O., Spyrtos, F., Harbeck, N., Dietrich, D. and et al. (2009) 'DNA methylation markers predict outcome in node-positive, estrogen receptor-positive breast cancer with adjuvant anthracycline-based chemotherapy', *Clin Cancer Res*, 15, pp. 315-323.
- Haupt, R. and Haupt, S. E. (2004) *Practical Genetic Algorithms*. 2nd edn. New York: Wiley-Interscience.
- Helzlsouer, K. J., Bush, T. L., Alberg, A. J. and et al. (1993) 'Prospective study of serum CA125 levels as markers of ovarian cancer', *Journal of the American Medical Association*, 269(1123-1126).
- Hintz, K. J. (1991) 'A measure of the information gain attributable to cueing', *IEEE Transactions on Man and Cybernetics Systems*, 21(2), pp. 434-442.

- Hiro, S., Komiya, Y. and Mizuta, M. (2002) 'Relative Projection Pursuit with an Extension of Friedman Index', *Proceedings of the 4th Conference of the Asian Regional Section of the International Association for Statistical Computing*, pp. 238-241.
- Hogdall, E. V., Nedergaard, S. L., Engelholm, S. A., Lundvall, L., Petri, A. L., Risum, S. and Hogdall, C. K. (2008) 'Novel biomarkers that predict survival in patients with ovarian cancer', *Journal of Clinical Oncology*, 26(15s).
- Holland, J. (1992) 'Genetic algorithms', *Scientific American*, 7, pp. 66-72.
- Hoyt, K., Warram, J. M. and et al. (2010) 'Determination of Breast Cancer Response to Bevacizumab Therapy Using Contrast-Enhanced Ultrasound and Artificial Neural Networks', *Journal of Ultrasound in Medicine*, 29(4).
- Hruschka, E. R., Castro, L. N. and Campello, R. J. G. B. (2004) 'Evolutionary algorithms for clustering gene-expression data', *Fourth IEEE International Conference on Data Mining*, pp. 403-406.
- Hsu, F. C., Kritchevsky, S., Liu, Y. and et al. (2009) 'Association between inflammatory components and physical function in the health, aging, and body composition study: a principal component analysis approach', *J Gerontol A Biol Sci Med Sc*, 64, pp. 581-589.
- Huang, Z. and Ng, M. K. (2003) 'A note on K-modes clustering', *Journal of Classification*, 20(2), pp. 257-261.
- Huber, P. J. (1981) *Robust Statistics*. London: John Wiley & Sons.
- Hyvarinen, A. (2001) 'Complexity Pursuit: Separating Interesting Components from Time Series', *Source Neural Computation*, 13(4), pp. 883-898.
- Jackson, J. E. (2003) *A User's Guide to Principal Components*. New York: WileyBlackwell.
- Jackson, W. C. and Norgard, J. D. (2008) 'A Hybrid Genetic Algorithm with Boltzmann Convergence Properties', *Journal of Optimization Theory and Applications*, 136(3), pp. 431-443.
- Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N. J., Chung, S., Emili, A., Snyder, M., Greenblatt, J. F. and Gerstein, M. (2003) 'A Bayesian networks approach for predicting protein-protein interactions from genomic data', *Science* 302, pp. 449-453.
- Jensen, F. V. (2001) *Bayesian networks and decision graphs*. New York: Springer.
- Jensen, R. and Shen, Q. (2004) 'Semantics-preserving dimensionality reduction: rough and fuzzy-rough approaches', *IEEE Transactions on Knowledge and Data Engineering*, 16(12), pp. 1457-1471.

- Jolliffe, I. T. (2002) *Principal Component Analysis*. 2nd edn. New York: Springer.
- Jones, G. (1998) 'Genetic and evolutionary algorithms', in Rague, P. (ed.) *Encyclopedia of Computational Chemistry*. New York: John Wiley and Sons.
- Joseph, G. M. and Patel, V. L. (1990) 'Domain knowledge and hypothesis generation in diagnostic reasoning', *Med. Decis Making*, 10, pp. 31-46.
- Kahn, C. E., Roberts, L. M., Shaffer, K. A. and Haddawy, P. (1997) 'Construction of a Bayesian network for mammographic diagnosis of breast cancer', *Computers in Biology & Medicine*, 27, pp. 19-29.
- Kamath, S. D. and Mahato, K. K. (2009) 'Principal component analysis (PCA)-based k-nearest neighbor (k-NN) analysis of colonic mucosal tissue fluorescence spectra', *Photomed Laser Surg*, 27, pp. 659-668.
- Kantardzic, M. (2002) *Data Mining: Concepts, Models, Methods, and Algorithms*. New Jersey: Wiley-IEEE Press.
- Karp, G. (2009) *Cell and Molecular Biology: concepts and experiments*. 6th edn. New York: John Wiley & Sons.
- Karzynski, M., Mateos, A. and et al. (2003) 'Using a Genetic Algorithm and a Perceptron for Feature Selection and Supervised Class Learning in DNA Microarray Data', *Artificial Intelligence* 20(2), pp. 39-51.
- Kawakami, S., Numao, N., Okubo, Y. and et al. (2008) 'Development, validation, and head-to-head comparison of logistic regression-based nomograms and artificial neural network models predicting prostate cancer on initial extended biopsy', *Eur Urol*, 54, pp. 601-611.
- Khan, J., Wei, J. S. and et al. (2001) 'Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks', *Nature Medicine*, 7(6), pp. 673-679.
- Kikkawa, F., Nawa, A., Kajiyama, H., Shibata, K., Ino, K. and Nomura, S. (2006) 'Clinical characteristics and prognosis of mucinous tumours of the ovary', *Gynecol Oncol*, 103, pp. 171-175.
- Kim, H., Golub, G. H. and Park, H. (2005) 'Missing value estimation for DNA microarray gene expression data: local least squares imputation', *Bioinformatics*, 21(2), pp. 187-198.
- Kohavi, R. A. (1995) 'Study of cross-validation and bootstrap for accuracy estimation and model selection', *In Proceedings of International Joint Conference on AI*, pp. 1137-1145.
- Koski, T. and Noble, J. (2009) *Bayesian Networks: An Introduction*. London: Wiley-Blackwell

- Kotsiantis, S. B., Kanellopoulos, D. and Pintelas, P. E. (2006) 'Data pre-processing for supervised learning', *International Journal of Computer Science*, 1(2), pp. 111-117.
- Krzanowski, W. J. and Hand, D. J. (2009) *ROC curves for continuous data*. London: Chapman and Hal.
- Krzysztof, S. and et al. (2007) 'An ant colony optimization algorithm for continuous optimization: application to feed-forward neural network training', *Neural Computing and Applications*, 16(3), pp. 235-247.
- Kulasingam, V., Pavlou, M. P. and Diamandis, E. P. (2010) 'Integrating high-throughput technologies in the quest for effective biomarkers for ovarian cancer', *Nat Rev Cancer*, 10, pp. 371-378.
- Lacave, C. and Diez, F. J. (2003) ' Knowledge acquisition in Prostanet, a Bayesian network for diagnosing prostate cancer', *Knowledge-Based Intelligent Information and Engineering Systems*, 2, pp. 1345-1350.
- Larma, J. and Gardner, G. J. (2006) 'Ovarian Cancer', in Bankowski, B. J. and et al. (eds.) *The Johns Hopkins Review of Gynecology and Obstetrics*. Philadelphia: Lippincott Williams and Wilkins, pp. 508-525.
- Larose, D. (2006) *Data Mining Methods and Models*. New Jersey: Wiley-IEEE Press.
- Le, T., Alshaikh, G. and et al. (2006) 'Prognostic Significance of Postoperative Morbidities in Patients With Advanced Epithelial Ovarian Cancer Treated With Neoadjuvant Chemotherapy and Delayed Primary Surgical Debulking', *Ann. Surg. Oncol*, 13, pp. 1711-1716.
- Ledermann, J. and et al. (2011) 'Phase II randomized placebo-controlled study of olaparib (AZD2281) in patients with platinum-sensitive relapsed serous ovarian cancer (PSR SOC)', *J Clin Oncol (Meeting Abstracts)*, 29.
- Lee, E. K., Cook, D., Klinke, S. and Lumley, T. (2005) 'Projection Pursuit for Exploratory Supervised Classification', *Journal of Computational and Graphical Statistics*, 14(4), pp. 831-846.
- Lee, J. K. (2010) *Statistical Bioinformatics: For Biomedical and Life Science Researchers*. New Jersey: Wiley-Blackwell.
- Lee, R. (2008) *Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing*. London: Springer.
- Leitao, M. M. and Barakat, R. R. (2009) 'Staging and Surgical Treatment', in Stack, M. S. and Fishman, D. A. (eds.) *Ovarian cancer*. 2nd edn. London: Springer.

- Leonard, T. and Hsu, J. S. J. (1999) *Bayesian methods: An analysis for statisticians and interdisciplinary researchers*. New York: Cambridge University Press.
- Lewis, S. and Menon, U. (2003) 'Screening for ovarian cancer', *Expert Review of Anticancer Therapy*, 3, pp. 55-62.
- Li, L., Weinberg, C. R., Darden, T. A. and Pedersen, L. G. (2001) 'Gene selection for sample classification based on gene expression data', *Bioinformatics*, 17(12), pp. 1131-1142.
- Lin, F. and He, G. (2005) 'An Improved Genetic Algorithm For Multi-Objective Optimization', *Sixth International Conference on Parallel and Distributed Computing*, pp. 938-940.
- Liu, C., Pan, C., Shen, J., Wang, H. and Yong, L. (2011) 'MALDI-TOF MS Combined With Magnetic Beads for Detecting Serum Protein Biomarkers and Establishment of Boosting Decision Tree Model for Diagnosis of Colorectal Cancer', *Int J Med Sci*, 8(1), pp. 39-47.
- Liu, D., Shi, T., DiDonato, J. A., Carpten, J. D., Zhu, J. and Duan, Z. (2004) 'Application of Genetic Algorithm/K-Nearest Neighbor Method to the Classification of Renal Cell Carcinoma', *Proceedings of the 2004 IEEE Computational Systems Bioinformatics Conference*, pp. 558-559.
- Liu, H., Hussain, F., Tan, C. L. and Dash, M. (2002) 'Discretization: An enabling technique', *Data Mining and Knowledge Discovery*, 6, pp. 393-423.
- Liu, H. and Motoda, H. (2007) *Computational Methods of Feature Selection*. New York: Chapman & Hall/CRC.
- Ma, P. C. H. and Chan, K. C. C. (2003) 'Discovering Clusters in Gene Expression Data Using Evolutionary Approach', *15th IEEE International Conference on Tools with Artificial Intelligence*.
- Maclin, P. S. and Dempsey, J. (1992) 'Using an artificial neural network to diagnose hepatic masses', *J Med Syst* 16(5), pp. 215-225.
- Maclin, P. S., Dempsey, J., Brooks, J. and Rand, J. (1991) 'Using neural networks to diagnose cancer', *J Med Syst*, 15(1), pp. 11-19.
- Mangasarian, O. L., Street, W. N. and Wolberg, W. H. (1995) 'Breast cancer diagnosis and prognosis via linear programming', *Operations Research*, 43(4), pp. 570-577.
- Marsland, S. (2009) *Machine Learning: An Algorithmic Perspective*. London: Chapman & Hall.
- Marzban, C. (2009) 'Basic statistics and basic AI: Neural Networks', in Haupt, S. E., Pasini, A. and Marzban, C. (eds.) *Artificial intelligence methods in the environmental sciences*. New York: Springer.

- Maurizi, M., Paludetti, G., Galli, J. and et al. (1999) 'Oncological and functional outcome of conservative surgery for primary supraglottic cancer', *Eur Arch Otorhinolaryngol*, 256, pp. 283-290.
- McLaren, C. E., Chen, W. P., Nie, K. and et al. (2009) 'Prediction of malignant breast lesions from MRI features: a comparison of artificial neural network and logistic regression techniques', *Acad Radiol*, 16, pp. 842-851.
- Mitchell, M. (1998) *An Introduction to Genetic Algorithms*. Cambridge: MIT Press.
- Mitra, S. and Acharya, T. (2003) *Data Mining: Multimedia, Soft Computing, and Bioinformatics*. New York: John Wiley.
- Molinaro, A. M., Simon, R. and Pfeiffer, R. M. (2005) 'Prediction error estimation: a comparison of re-sampling methods', *Bioinformatics*, 21, pp. 3301-3307.
- Mukherjee, S., Tamayo, P., Mesirov, J. P., Slonim, D., Verri, A. and Poggio, T. (1999) 'Support vector machine classification of microarray data', *Technical Report MIT*, 182.
- Mukhopadhyay, A. and al., E. (2010) 'Development of a functional assay for homologous recombination status in primary cultures of epithelial ovarian tumor and correlation with sensitivity to poly(ADP-ribose) polymerase inhibitors', *Clinical Cancer Research*, 16(8), pp. 2344-2351
- Munakata, T. (1998) *Fundamentals of the New Artificial Intelligence*. New York: Springer.
- Muyldermans, M., Cornillie, F. J. and Koninckx, P. R. (1995) 'CA125 and endometriosis', *Hum Reprod Update*, 1(2), pp. 173-187.
- Naguib, R. N. G. and Sherbet, G. V. (2001) *Artificial Neural Networks in Cancer Diagnosis, Prognosis, and Patient Management*. New York: CRC Press.
- Narayanan, A., Keedwell, E. C. and Olsson, B. (2002) 'Artificial intelligence techniques for bioinformatics', *Appl. Bioinformatics* 1, pp. 191-222.
- Neapolitan, R. E. (2004) *Learning Bayesian Networks*. New Jersey Prentice-Hall.
- Nicolini, C., Gaglio, S. and Ruggiero, C. (1989) 'Artificial intelligence techniques for the control of cancer cells', *Cell Biochemistry and Biophysics* 14(2), pp. 117-127.
- Niedermayer, D. (1998) *An Introduction to Bayesian Networks and their Contemporary Applications*. Available at: <http://www.niedermayer.ca/papers/bayesian/bayes.html> (Accessed: 20-03-2009).

- O'Neill, E. S., Dluhy, N. C. and Chun, E. (2005) 'Modelling novice clinical reasoning for a computerised decision support system', *Advanced Nursing*, 49(1).
- Olson, D. L. and Delen, D. (2008) *Advanced Data Mining Techniques*. Berlin Springer.
- Patil, B. M., Joshi, R. C. and Toshniwal, D. (2010) 'Impact of K-Means on the Performance of Classifiers for Labelled Data', in Ranka, S., Banerjee, A., Biswas, K. K., Dua, S., Mishra, P., Moona, R. and al., E. (eds.) *Contemporary Computing: Second International Conference*. Berlin: Springer, pp. 423-434.
- Patnaik, S. K., Kannisto, E., Knudsen, S. and Yendamuri, S. (2010) 'Evaluation of microRNA expression profiles that may predict recurrence of localized stage I non-small cell lung cancer after surgical resection', *Cancer Res*, 70, pp. 36-45.
- Pena-Reyes, C. A. and Sipper, M. (2000) 'A fuzzy genetic approach to breast cancer diagnosis', *Artificial Intelligence in Medicine*, 17, pp. 131-155.
- Petricoin, E. F., Ornstein, D. K., Paweletz, C. P. and et al. (2002) 'Serum proteomic patterns for detection of prostate cancer', *Natl Cancer Inst*, 94, pp. 1576-1578.
- Plante, M., Lau, S., Brydon, L., Swenerton, K., LeBlanc, R. and Roy, M. (2006) 'Neoadjuvant chemotherapy followed by vaginal radical trachelectomy in bulky stage IB1 cervical cancer: case report', *Gynecol Oncol*, 101, pp. 367-370.
- Press, W. H., Flannery, B. P., Teukolsky, S. A. and Vetterling, W. T. (1992) *Numerical Recipes in FORTRAN: The Art of Scientific Computing*. Cambridge: Cambridge University Press.
- Provost, F., Fawcett, T. and Kohavi, R. (1998) 'The case against accuracy estimation for comparing induction algorithms', *Proceedings of the 15th International Conference on Machine Learning*, 15, pp. 445-453.
- Quinlan, J. R. (1986) 'Induction of decision trees', *Machine Learning*, 1, pp. 81-106.
- Quinlan, J. R. (1993) *C4.5: Programs for Machine Learning*. 1st edn. London: Morgan Kaufmann.
- Reeves, C. R. and Rowe, J. E. (2003) *Genetic Algorithms: Principles and Perspectives*. New York: Kluwer Academic Publishers.
- Resnick, K. E., Alder, H., Hagan, J. P., Richardson, D. L., Croce, C. M. and Cohn, D. E. (2009) 'The detection of differentially expressed microRNAs from the serum of ovarian cancer patients using a novel real-time PCR platform', *Gynecol Oncol*, 112, pp. 55-59.

- Riffenburgh, R. H. (2005) *Statistics in medicine*. 2nd edn. London: Academic Press.
- Rodriguez, G. C., Soper, J. T. and et al. (1992) 'Improved palliation of cerebral metastases in epithelial ovarian cancer using a combined modality approach including radiation therapy, chemotherapy and surgery', *Clin Oncol*, 10, pp. 1553-1560.
- Rodríguez-Piñeiro, A. M., Rodríguez-Berrocal, F. J. and Páez de la Cadena, M. (2007) 'Improvements in the search for potential biomarkers by proteomics: application of principal component and discriminant analyses for two-dimensional maps evaluation', *Chromatogr B Analyt Technol Biomed Life Sci*, 849, pp. 251-260.
- Rokach, L. and Maimon, O. (2008) *Data Mining with Decision Trees: Theory and Applications*. London: World Scientific Publishing Company.
- Rothenberg, M. L., Liu, P. Y. and et al. (2003) 'Combined Intraperitoneal and Intravenous Chemotherapy for Women With Optimally Debulked Ovarian Cancer: Results From an Intergroup Phase II Trial', *Clin Oncol*, 21, pp. 1313-1319.
- Royston, P., Altman, D. G. and Sauerbrei, W. (2006) 'Dichotomizing continuous predictors in multiple regressions: a bad idea', *Stat Med*, 25, pp. 127-141.
- Russel, S. J. and Norvig, P. (1995) *Artificial Intelligence: A Modern Approach*. New Jersey: Prentice Hall.
- Saeyns, Y. and et al. (2007) 'A review of feature selection techniques in bioinformatics', *Bioinformatics*, 23(19), pp. 2507-2517.
- Salzberg, S. (1997) 'On comparing classifiers: pitfalls to avoid and a recommended approach', *Data Min. Knowl. Disc.*, 1(3), pp. 317-328.
- Scherf, U., Ross, D. T. and et al. (2000) 'A Gene Expression Database for the Molecular Pharmacology of Cancer', *Nature Genetics*, 24(3), pp. 236-244.
- Scholkopf, B., Smola, A. J. and Müller, K. R. (1998) 'Nonlinear component analysis as a kernel eigenvalue problem', *Neural Computation*, 10, pp. 1299-1319.
- Scholz, M. (2006) *Approaches to analyze and interpret biological profile data*. POTSDA [Online]. Available at: http://opus.kobv.de/ubp/volltexte/2006/783/pdf/scholz_diss.pdf.
- Seidman, J. D. and Kurman, R. J. (2003) 'Pathology of ovarian carcinoma', *Hematol Oncol Clin North Am*, 17, pp. 909-925.
- Sekino, M. and Nitta, K. (2007) 'Unbiased likelihood backpropagation learning', in Ishikawa, M., Doya, K., Miyamoto, H. and Yamakawa, T. (eds.) *Neural Information Processing: 14th International Conference, ICONIP 2007*. Berlin: Springer.

- Seltzer, V. L. (1999) 'Ovarian cancer', in Karmar, B. S. and et al. (eds.) *Cancer screening: Theory and practice*. New York: Marcel Dekker.
- Setiono, R. (1996) 'Extracting rules from pruned neural networks for breast cancer diagnosis', *Artif. Intell. Med.* , 8, pp. 37-51.
- Simon, M. A. (2005) 'Ovarian cancer', in Smith, D. S., Sullivan, L. E. and Hay, S. F. (eds.) *Field Guide to Internal Medicine*. Philadelphia: Lippincott Williams & Wilkins.
- Simon, R., Radmacher, M. D., Dobbin, K. and McShane, L. M. (2003) 'Pitfalls in the use of DNA microarray data for diagnostic and prognostic classification', *J Natl. Cancer Inst.*, 95, pp. 14-18.
- Skala, M., Rosewall, T., Dawson, L. and et al. (2007) 'Patient-assessed late toxicity rates and principal component analysis after image-guided radiation therapy for prostate cancer', *Int J Radiat Oncol Biol Phys*, 68, pp. 690-698.
- Skates, S. J., Xu, F. J., Yu, Y. H. and et al. (1995) 'Toward an optimal algorithm for ovarian cancer screening with longitudinal tumor markers', *Cancer*, 76, pp. 2004-2010.
- Slade, P. and Gedeon, T. D. (1993) 'Bimodal Distribution Removal', in Mira, J., Cabestany, J. and Prieto, A. (eds.) *Proceedings of the International Workshop on Artificial Neural Networks*. Berlin: Springer.
- Smith, E. R., Qi Cai, K., Capo-chichi, C. D. and Xi Xu, X. (2009) 'Aberrant Epithelial Differentiation in Ovarian Cancer', in Stack, M. S. and Fishman, D. A. (eds.) *Ovarian cancer*. 2nd edn. London: Springer.
- Sood, A. K. and Gershenson, D. M. (2005) 'Management of early-stage ovarian cancer', in Bristow, R. E. and Karlan, B. Y. (eds.) *Surgery for Ovarian Cancer: Principles and Practice*. Oxon: Taylor & Francis.
- Sotoca, J. M., Pla, F. and Sanchez, J. S. (2007) 'Band Selection in Multispectral Images by Minimization of Dependent Information Systems, Man, and Cybernetics, Part C: Applications and Reviews', *IEEE Transactions on Computers*, 37(2), pp. 258-267.
- Speight, P. M., Elliot, A. E., Jullien, J. A., Downer, M. C. and Zakzrewska, J. M. (1995) 'The use of artificial intelligence to identify people at risk of oral cancer and precancer', *Br Dent J* 179, pp. 382-387.
- Spiegel, M. R. (1992) *Theory and Problems of Probability and Statistics*. 2nd edn. New York: McGraw-Hill.
- Stack, M. S. and Fishman, P. A. (2009) *Ovarian cancer*. 2nd edn. London: Springer.

- Stephan, C., Buker, N., Cammann, H., Meyer, H. A. and et al. (2008) 'Artificial neural network (ANN) velocity better identifies benign prostatic hyperplasia but not prostate cancer compared with PSA velocity', *BMC Urol*, 8(10).
- Su, F., Lang, J., Kumar, A., Hsieh, B. and Marc, A. (2007) 'Validation of Candidate Serum Ovarian Cancer Biomarkers for Early Detection', *Biomarker Insights*, 2, pp. 369-375.
- Golub, T. R. and et al. (1999) 'Molecular classification of cancer: class discovery and class prediction by gene expression monitoring', *Science*, 286(5439), pp. 531-537.
- Teramukai, S. and et al. (2007) 'PIEPOC: A New Prognostic Index for Advanced Epithelial Ovarian Cancer Japan Multinational Trial Organization OC01-01 10.1200/JCO.2007', *J Clin Oncol.*, 25(22), pp. 3302-3306.
- Thackery, E. (2005) *Gale encyclopedia of cancer*. 2nd edn. Detroit: Gale Group.
- Thompson, C. and Dowding, C. (2002) *Clinical Decision Making and Judgement in Nursing*. London: Churchill Livingstone.
- Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. and Altman, R. B. (2001) 'Missing value estimation methods for DNA microarrays', *Bioinformatics*, 17(6), pp. 520-525.
- Van der Gaag, L. C., Renooij, S., Witteman, C. L. M., Aleman, B. M. P. and Taal, B. G. (2002) 'Probabilities for a probabilistic network: a case study in oesophageal cancer', *Artificial Intelligence in Medicine*, 25(2), pp. 123-148.
- Vasey, P. A., Jayson, G. C., Gordon, A., Gabra, H., Coleman, R., Atkinson, R., Parkin, D., Paul, J., Hay, A. and Kaye, S. B. (2004) 'Phase III Randomized Trial of Docetaxel-Carboplatin Versus Paclitaxel-Carboplatin Versus Paclitaxel-Carboplatin as First-line Chemotherapy for Ovarian Carcinoma', *JNCI J Natl Cancer Inst*, 96, pp. 1682-1691.
- Venkitaraman, A. R. (2002) 'Cancer Susceptibility and the Functions of BRCA1 and BRCA2', *Cell* 108(2), pp. 171-182.
- Vergote, I., Trope, C. G., Amant, F., Kristensen, G. B., Ehlen, T., Johnson, N. and et al. (2010) 'Neoadjuvant chemotherapy or primary surgery in stage IIIC or IV ovarian cancer', *New England J Med*, 363, pp. 943-953.
- Vergote, I., Van Gorp, T., Amant, F., Neven, P. and Berteloot, P. (2005) 'Neoadjuvant chemotherapy or ovarian cancer', *Oncology* 19, pp. 1615-1622.
- Vlahou, A., Schorge, J. O., Gregory, B. W. and Coleman, R. L. (2003) 'Diagnosis of ovarian cancer using decision tree classification of mass spectral data', *J. Biomed. Biotechnol.*, 10, pp. 308-314.

- Wang, X., Zheng, B., Good, W., King, J. and Chang, Y. (1999) 'Computer assisted diagnosis of breast cancer using a data-driven Bayesian belief network', *International Journal of Medical Informatics*, 54, pp. 115-126.
- Wang, Y., Shi, Y., Yue, B. and Teng, H. (2010) 'An efficient differential evolution algorithm with approximate fitness functions using neural networks', in Deng, H. and Lei, J. (eds.) *Artificial Intelligence and Computational Intelligence: International Conference, AICI 2010*. Berlin: Springer.
- Wilding, P., Morgan, M. A., Grygotis, A. E., Shoffner, M. A. and Rosato, E. F. (1994) 'Application of backpropagation neural networks to diagnosis of breast and ovarian-cancer', *Cancer Letters*, 77, pp. 145-153.
- Wilkinson, S. J. and et al. (2008) 'Expression of gonadotrophin releasing hormone receptor I is a favorable prognostic factor in epithelial ovarian cancer', *Human Pathology*, 39(8), pp. 1197-1204.
- Williams, R. M., Flesken-Nikitin, A., Ellenson, L. H., Connolly, D. C., Hamilton, T. C., Nikitin, A. Y. and Zipfel, W. R. (2010) 'Strategies for high-resolution imaging of epithelial ovarian cancer by laparoscopic nonlinear microscopy', *Transl Oncol*, 3(3), pp. 181-194.
- Windsor, J. A., Knight, G. S. and Hill, G. L. (1998) 'Wound healing response in surgical patients: recent food intake is more important than nutritional status', *Br J Surg.*, 75(2), pp. 135-137.
- Wooster, R. and Weber, B. L. (2003) 'Breast and ovarian cancer', *N Engl J Med*, 348, pp. 2339-2347.
- Wulfkuhle, J. D., Liotta, L. A. and Petricoin, E. F. (2003) 'Proteomic applications for the early detection of cancer', *Nature Rev. Cancer*, 3, pp. 267-275.
- Yang, Y. and Pedersen, J. (1997) 'A comparative study on feature selection in text categorization', *In Proceedings ICML-97*, pp. 412-420.
- Yao, I., Sugiura, Y., Matsumoto, M. and Setou, M. (2008) 'In situ proteomics with imaging mass spectrometry and principal component analysis in the scrapper-knockout mouse brain', *Proteomics*, 8, pp. 3692-3701.
- Yap, T. A., Kaye, S., Ashworth, A. and Tutt, A. (2010) 'Tumour-Specific Synthetic Lethality: Targeting BRCA Dysfunction in Ovarian Cancer', in Kate, S., Brown, R., Gabra, H. and Gore, M. (eds.) *Emerging Therapeutic Targets in Ovarian Cancer*. London: Springer.
- Yeung, K. Y., Haynor, D. R. and W.L., R. (2001) 'Validating Clustering for Gene Expression Data', *Bioinformatics* 17(4), pp. 309-318.
- Young, C. (1987) 'Intuition and nursing process', *Holistic Nursing Practice*, 1(3), pp. 52-62.
- Zaknich, A. (2003) *Neural networks for intelligent signal processing*. London: World Scientific Publishing Co Pte Ltd

- Zhang, Z., Bast, R. C., Yu, Y. and et al. (2004) 'Three biomarkers identified from serum proteomic analysis for the detection of early stage ovarian cancer', *Cancer Res.*, 64(16), pp. 5882-5890.
- Zhou, Y., Lu, Y. and Shi, C. (1997) 'Combining neural network, genetic algorithm and symbolic learning approach to discover knowledge from databases', *In Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics*, 7, pp. 4388-4393.
- Zhou, Z. H., Jiang, Y., Yang, Y. B. and Chen, S. F. (2002) 'Lung cancer cell identification based on artificial neural network ensembles', *Artificial Intelligence in Medicine*, 24(1), pp. 25-36.
- Zhu, Z., Ong, Y. and Dash, M. (2007) 'Markov blanket-embedded genetic algorithm for gene selection', *Pattern Recogn*, 40(11), pp. 3236-3248.
- Zurawski, V. R. J., Orjaseter, H., Andersen, A. and et al. (1988) 'Elevated serum CA125 levels prior to diagnosis of ovarian neoplasia: relevance for early detection of ovarian cancer', *Int. J. Cancer*, 42, pp. 677-680.

Final page