NEWCASTLE UNIVERSITY

A Reliable Neural Network-Based Decision Support System for Breast Cancer Prediction

Shirin Ameiryan Mojarad

Doctor of Philosophy

a thesis is submitted to the School of Electrical, Electronic and Computer Engineering, Newcastle University, United Kingdom, in partial fulfilment of the requirements for the degree of Doctor of Philosophy in Electrical and Electronic Engineering

April 2012

Abstract

Axillary lymph node (ALN) metastasis status is an important prognostic marker in breast cancer and is widely employed for tumour staging and defining an adjuvant therapy. In an attempt to avoid invasive procedures which are currently employed for the diagnosis of nodal metastasis, several markers have been identified and tested for the prediction of ALN metastasis status in recent years. However, the nonlinear and complex relationship between these markers and nodal status has inhibited the effectiveness of conventional statistical methods as classification tools for diagnosing metastasis to ALNs. The aim of this study is to propose a reliable artificial neural network (ANN) based decision support system for ALN metastasis status prediction. ANNs have been chosen in this study for their special characteristics including nonlinear modelling, robustness to inter-class variability and having adaptable weights which makes them suitable for data driven analysis without making any prior assumptions about the underlying data distributions. To achieve this aim, the probabilistic neural network (PNN) evaluated with the .632 bootstrap is investigated and proposed as an effective and reliable tool for prediction of ALN metastasis. For this purpose, results are compared with the multilayer perceptron (MLP) neural network and two network evaluation methods: holdout and cross validation (CV). A set of six markers have been identified and analysed in detail for this purpose. These markers include tumour size, oestrogen receptor (ER), progesterone receptor (PR), p53, Ki-67 and age. The outcome of each patient is defined as metastasis or non-metastasis, diagnosed by surgery. This study makes three contributions: firstly it suggests the application of the PNN as a classifier for predicting the ALN metastasis, secondly it proposes a the .632 bootstrap evaluation of the ANN outcome, as a reliable tool for the purpose of ALN status prediction, and thirdly it proposes a novel set of markers for accurately predicting the state of nodal metastasis in breast cancer. Results reveal that PNN provides better sensitivity, specificity and accuracy in most marker combinations compared to MLP. The results of evaluation methods' comparison demonstrate the high variability and the existence of outliers when using the holdout and 5-fold CV methods. This variability is reduced when using the .632 bootstrap. The best prediction accuracy, obtained by combining ER, p53, Ki-67 and age was 69% while tumour size and p53 were the most significant individual markers. The classification accuracy of this panel of markers emphasises their potential for predicting nodal spread in individual patients. This approach could significantly reduce the need for invasive procedures, and reduce post-operative stress and morbidity. Moreover, it can reduce the time lag between investigation and decision making in patient management.

To my beloved parents and brother

Acknowledgements

I would like to thank my supervisors Prof. Satnam S. Dlay and Dr Wai Lok Woo for all their guidance, support and patience. I have been privileged to work under their supervision, and I truly appreciate their infinite help and encouragement. I am also grateful to Dr. Gajanan V. Sherbet for his help and support for the medical part of the study. And special thanks to Mrs. Gillian Webber. I am very grateful for her kindness, administrative support and help.

I want to express acknowledgement to the ORS Award Scheme committee, the School of Electrical, Electronic and Computer Engineering and Newcastle University for their generous financial support through numerous awards and scholarships.

I would like to thank all my friends who made my PhD journey a pleasant one, those who are in my heart and not on paper. I would like to express my gratitude to Reza, Mahsa, Sadaf, Juila, Kamyar, Gholamreza and Mohammad for helping me with the details that only a sharp eye could spot.

I am grateful to my Aunt and Uncle, Tahereh and Reza, for their kindness and constant support during my studies. Last but not least, I would like to thank my beloved parents and brother for their never-fading love, support and encouragement. I dedicate this thesis to them.

Contents

List of figures

List of tables

Nomenclature

$PCC_{A,B}$	Pearson correlation coefficient between variables A and B
$cov(A,B)$	Covariance of variables A and B
$E(.)$	Expected value
S_A	Standard deviation of vector A
μ_A	Mean of vector A
X	Input vector $X = \{x_1, x_2, \ldots, x_N\}$ with $i = 1{:}N$ samples
$\mathrm{X_p}$	Input matrix $X_p = \{X_1, X_2, \ldots, X_P\}$ with $p = 1{:}P$ input dimension and $X_p = \{x_1, x_2, \ldots, x_N\}$ with $i = 1{:}N$ samples
T	Target vector $T = \{t_1, t_2, \ldots, t_N\}$ with $i = 1{:}N$ samples
(x_i, t_i)	A data point with the input x_i and target value t_i
N	Sample size (number of patients)
$\mathrm{N_c}$	Number of samples in class c, For a dichotomous outcome with $c = 0{,}1$, N_1 and N_2 represent the sample size in each class
e	Error defined as the difference between the computed value and the desired target
β	Regression coefficient
$\bar{x}$	Mean of vector x
$\bar{t}$	Mean of vector t
$\pi(x)$	Conditional mean of the output T given the input x
μ	Mean
$\tilde{\mu}$	Mean of transferred data in LDA
$\tilde{S}$	Standard deviation of transferred data in LDA
w	Transformation vector in LDA / weights of artificial neural network
w^T	Transpose of vector w
$C(w)$	Cost function to be maximized in LDA

s_B	between-class scatter matrix in LDA
S_W	within-class scatter matrix in LDA
P_c	proportion of the data points belonging to class c
V	Input variable
$Values(V)$	Possible values of input variable V
X_v	the subset of vector X for which variable V has values v
$\varphi(.)$	Transfer function
s	s^{th} iteration (step) of the training process
β	the learning rate (step size) in back propagation
E_i	output error when testing input i
o_{jp}	predicted output of j^{th} output neuron for p^{th} input pattern
w_{hj}	Weights connecting h^{th} hidden node to j^{th} output neuron
y_h	vector of outputs of the hidden layer h
ξ	MSE cost function
v_j	Activation of neuron j
δ	local gradient vector
W	Wight matrix
H	Hessian matrix
w_s	Multilayer perceptron weight at step s
d_s	search direction at step s
g_s	gradient vector of the error surface at step s
$\widetilde{H}$	modified value of hessian matrix H
I	unit matrix
α	A positive coefficient added to hessian matrix H to ensure it is positive
$P(O = j\|X)$	Conditional probability of $O = j$ given X
$f_j(X)$	likelihood function equal to class-conditional PDF $P(X\|O = j)$

π_j	prior probability of class *j*
$k(.)$	Kernel function
E^{a}	Apparent error
E^{b}	Bootstrap error
$E^{.632b}$	.632 bootstrap error
AUC	Area under curve
$\widehat{AUC}$	Area under curve predicted from samples
$\mathrm{x}^{*\mathrm{b}}$	*b*th bootstrap sample
$\overline{\mathrm{x}^{*\mathrm{b}}}$	Data set including the samples from x left out of $\mathrm{x}^{*\mathrm{b}}$
B	Total number of bootstrap samples *b = 1:B*

Abbreviations

ALN	Axillary Lymph Node
ANN	Artificial Neural Network
AUC	Area Under Curve
BP	Back Propagation
DA	Discriminant Analysis
DSS	Decision Support Systems
DT	Decision Tree
ER	Oestrogen Receptor
FN	False Negative
FNA	Fine Needle Aspirate
FP	False Positive
FPR	False Positive Rate
GD	Gradient Descent
IHC	Immunohistochemistry
LDA	Linear Discriminant Analysis
LR	Logistic Regression
ML	Maximum Likelihood
MLP	Multilayer Perceptron
MSE	Mean Square Error
PCC	Pearson’s Correlation Coefficient
PDF	Probability Density Function
PNN	Probabilistic Neural Network
PR	Progesterone Receptor
RBF	Radial Basis Function
ROC	Receiver Operating Characteristic
SCG	Scaled Conjugate Gradient
SLNB	Sentinel Lymph Node Biopsy
TN	True Negative
TP	True Positive
TPR	True Positive Rate

Awards

Upon the exceptionally successful completion of my master degree as the top student with the overall of 84% and obtaining the highest mark for my research dissertation amongst all students, I received three scholarships from different sources to pursue my PhD research:

- Overseas Research Students Awards Scheme (£28000)
- EECE scholarship for outstanding students, Newcastle University (£21000)
- Newcastle University International Postgraduate Scholarship (£4500)

Awarded the best poster prize for the SSC group in the PG conference 2009, EECE, Newcastle University.

National Instruments award for best paper and presentation for the SSC group in PG conference 2010, EECE, Newcastle University.

Publications

[1] S. A. Mojarad, S. S. Dlay, W. L. Woo, and G. V. Sherbet, "Reliable Prediction of Breast Cancer Progression and Prognosis Using ANN- Based Analysis of a Class of Minimally Invasive Biomarkers," *Submitted in IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*.

[2] S. A. Mojarad, S. S. Dlay, W. L. Woo, and G. V. Sherbet, "Cross Validation Evaluation for Breast Cancer Prediction Using Multilayer Perceptron Neural Networks," *The American Journal of Engineering and Applied Sciences* vol. 4, pp. 576-585, 2012.

[3] S. A. Mojarad, S. S. Dlay, W. L. Woo, and G. V. Sherbet, "Breast cancer prediction and cross validation using multilayer perceptron neural networks," in *7th IEEE, IET International Symposium on Communication Systems Networks and Digital Signal Processing (CSNDSP),* 2010, pp. 760-764.

[4] S. A. Mojarad, S. S. Dlay, W. L. Woo, and G. V. Sherbet, "An artificial neural network-based evaluation of tumour biomarkers for the prediction of nodal spread and prognosis of breast cancer," in *The 8th International Conference of Anticancer Research* Kos, Greece, 2008.

Chapter 1

1. Introduction

1.1. Motivation

Breast cancer is the major cause of cancer death amongst women and it has been identified as the most prevalent cancer type amongst women in England in 2009 which afflicted 40,260 new women [1]. There has been a sharp decrease in breast cancer mortality in recent years due to the application of early surgery and systemic therapy such as chemotherapy or hormonal therapy for disease management [2]. However, besides reducing the mortality rate, these treatments cause various side effects such as patient morbidity and are a tremendous cost to treatment centres. In order to reduce the patient morbidity and treatment costs as well as mortality, it is important to both avoid unnecessary treatment and not to miss cancer-positive cases. Diagnosis of cancer-positive patients is especially important as successful cancer management through adjuvant therapies[1] and surgery is significantly improved by early diagnosis.

Early diagnosis, identifying patients who need treatment and defining the adjuvant therapy depends on many factors including axillary lymph nodes (ALNs) metastasis (cancer cells present in the ALNs), tumour size, menopausal status, presence of hormone receptors and abnormal production of certain proteins such as HER2/neu in tumour. Amongst these factors, the presence of tumour in the ALNs and the corresponding number of involved nodes are regarded as the most important factors affecting the medical decision made for adjuvant therapy and surgery in patients with breast carcinoma [3-5].

For many years, the presence of ALN metastasis has been diagnosed by axillary dissection. Nevertheless, dissection is associated with marked morbidity and is unnecessary for patients with negative ALN metastasis. Sentinel lymph node biopsy (SLNB) is currently practiced widely in the United Kingdom as an alternative approach to axillary dissection [6]. SLNB is considered to be an accurate and reliable method for ALN sampling which causes less morbidity than axillary dissection, especially for women with node-negative breast cancer. However, SLNB is still an invasive procedure with its associated complications such as pain and paraesthesia [7].

[1] Adjuvant therapy in cancer refers to the treatment which is given in addition to surgery. Adjuvant therapies, such as chemotherapy, are advised where the initial tumour has been removed but there is still a statistical risk of relapse. In Latin "adjuvans" means to help and, particularly, to help reach a goal.

Fisher et al. have shown that identifying an adjuvant choice of treatment and disease management is highly dependent on the state of tumour dissemination and the chance of survival [8]. Although accurate and reliable detection of metastasis in the ALNs is viable through axillary dissection, an alternative non-invasive method for identifying the state of tumour dissemination has been keenly sought.

Since there has been an observation of microscopic metastasis in early-stage breast cancer in a significant number of patients, novel prognostic markers have become the area of interest for breast cancer prognosis. Many studies, utilizing specialized markers of tumour, have proven that these markers can be utilized as an alternative to dissection for ALN metastasis prediction [9-11]. A large variety of these biomarkers and their prognostic value in different kinds of cancer have been explored to make an accurate prognosis about tumour progression to help physicians identify the best therapy for the disease.

Many cellular and molecular markers have been identified to be clinically useful for early diagnosis of breast cancer. However, the large number of available biomarkers and the complex relation of these markers with the state of tumour progression have negated the efforts of medical experts for accurate breast cancer prognosis based on these markers. Hence, computational techniques have been attempted for an accurate and reliable prognosis based on these factors.

1.2. Problem statement

Although axillary dissection and SLNB offer an accurate and reliable detection of ALN metastasis, a less invasive method for identifying the state of tumour dissemination is more favourable for both patients and medical centres in terms of cost, mortality and morbidity. Several tumour biomarkers have been utilized as an alternative to dissection for ALN metastasis prediction [9-11]. Nevertheless, there is little agreement to date on a single or a combination of predictive markers for accurate nodal status prediction that can replace the conventional techniques such as SLNB [7]. Large number of available tumour characteristics, the nonlinear and complex relationship between markers and nodal involvement and small sample sizes available for analysis are the main issues hindering ALN prediction with tumour characteristics as an alternative to the invasive procedures.

1.2.1. Abundance and complex nature of biomarkers

Many clinical, pathological and molecular markers have been identified to be clinically useful for early diagnosis of breast cancer. However, the large number of available biomarkers and the complex relation of these markers with the state of tumour progression have negated the efforts of medical experts for accurate breast cancer prognosis based on these markers.

In addition, many new prognostic factors demonstrate non-monotonic characteristics i.e. not constantly increasing or decreasing. The prognostic value of some markers also depend on their interaction with other markers. Therefore it would be desirable to identify the prognostic value of the potential prognostic markers. Seker et al. [12] have employed a multilayer perceptron (MLP) for identifying the most and least significant prognostic factors for breast cancer survival analysis by extracting feature evaluation indices. However, they confirm a level of uncertainty on the obtained ranking of features.

1.2.2. Data availability

One major issue in cancer prediction using tumour characteristics is that data sets relating to cancer typically can contain a small number of patients in comparison with the number of measured variables. This is often due to the high cost of extracting, i.e. detecting and measuring these markers and establishing the relationship between degrees of expression and disease progression over a period of time. This limitation results in two problems in classification. First, the classifier will learn the existing patterns in the population using only a small number of samples. Second, an estimate of the accuracy of the classifier for new data can only be performed using the same small data set.

1.3. Computer aided diagnosis

To address the above issues, a wide range of studies have investigated various statistical and artificial intelligence methods for the prediction of nodal involvement, exploiting the ever-increasing number of biomarkers deemed to be effective in disease dissemination and metastasis.

Statistical methods have been conventionally applied to the prognosis and classification of breast cancer. However, the inefficiency of these techniques in analysing the nonlinear interaction of tumour biomarkers has initiated the necessity for more accurate and complex classification models. This inefficiency originates from the intrinsic characteristic of statistical methods such as assumptions about underlying distribution of data and requiring large sample size in each output class [13].

An alternative method which has been successfully employed in the field of medical diagnosis and prognosis is artificial neural network (ANN). It has been proven that ANN results in cancer detection and prediction are better or at least comparable to the results obtained from statistical methods. Burke et al. [14] have shown that the predictive accuracy of ANN for five-year survival prediction is significantly better than a statistical method such as principal components analysis (PCA), classification and regression trees and logistic regression (LR).

It is reported that artificial neural networks (ANN) provide a powerful method of analysing the inherently complex nature of potential cancer proliferation markers [15, 16]. In breast cancer, ANN has proved to be a reliable tool for predicting metastasis to the ALNs, using tumour's cellular markers [17]. Therefore, ANN's ability to identify highly nonlinear relationships between markers makes it a versatile tool in evaluating the dissemination of tumour proliferation data. Previous studies have shown that ANNs outperform most traditional and modern statistical techniques such as LR [18] and discriminant analysis (DA) [19].

Among various ANN architectures, MLP has been the most widely used method for cancer prediction and prognosis [20-22]. The effectiveness of MLPs in breast cancer diagnosis has been evaluated using clinical, pathological and immunohistochemical data which recommend MLP as a powerful technique for cancer prediction [21]. Nevertheless, MLPs suffer from some common drawbacks including network complexity, and numerous network variables require random weight initialisation and have to be optimised.

Another ANN architecture that has been found suitable for numerous applications in recent years is the probabilistic neural network (PNN). The PNN, first introduced by Specht in 1988 and 1990, is capable of defining decision boundaries exploiting Bayesian theory [23, 24]. This quality makes the PNN an effective technique for classification and pattern recognition tasks, especially when used for data sets containing samples adequately correlated to the associated class and well separated categories [25]. The PNN also trains faster than the MLP and requires fewer variables to be optimised. Although this method has been known for some time, it has only recently been identified as an effective classifier for cancer diagnosis and prediction [26, 27].

1.4. Aims and objectives

The primary aim in this study is to make an original contribution to the development of neural network models for breast cancer prognosis in terms of predicting the state of ALNs involvement. The biomolecular data selection for predicting the nodal involvement status requires an accurate system. Therefore, the second aim of the project is to choose a reliable error estimation method for the ANNs employed in this study.

A further objective is to accurately identify the contribution of each biomarker considered in this study to the prediction process. This is demonstrated by using the designed ANN to determine the potential of these markers and the most reliable prediction from the least costly set of markers. Thus the objectives are mapped out as follows:

- to introduce a specifically designed ANN for predicting the state of nodal involvement in breast cancer;
- to develop a reliable analysis of the estimation of the designed ANN error rate, which can offer more reliable error estimation for small sample size in breast cancer prediction;
- to gain insight into the importance of breast cancer biomarkers and to define an original effective combination of patient information and tumour markers for predicting ALN metastasis in breast cancer.

1.5. Methodology and contributions

In this study, the PNN is proposed as a classification platform to predict ALN metastasis in a group of patients suffering from breast cancer. The MLP is employed as a benchmark method to be compared against the PNN due to its frequent application in breast cancer prognosis. The results of the PNN are reported and compared against the results obtained from the MLP in terms of accuracy, sensitivity and specificity. Using two ANN architectures, it is intended to investigate various ways of using neural network models to extend traditional statistical models for cancer prognosis. These models are expected to provide better results than linear statistical methods as they would be able to model both non-linear effects of prognostic factors and interactions between them.

Another aim of the study is to choose a reliable error estimation method for the ANNs employed in this research. For this purpose, the variance of predictions obtained by .632 bootstrap, 5-fold cross validation and holdout methods are computed and compared.

To achieve insight into the importance of breast cancer biomarkers, a novel set of six biomarkers is proposed for the purpose of predicting the progression of breast carcinoma. While the relationship between some of these markers and nodal metastasis is well established in medical literature, no information is available on the degree of effectiveness of combining these features for ALN metastasis prediction in breast cancer. These biomarkers are analysed for their predictive significance associated with metastasis to ALNs and include tumour size together with the following five markers: oestrogen receptor (ER), progesterone receptor (PR), p53, Ki-67 expression and age. The data relating to these markers has been obtained using minimally invasive procedures in which a sample of tumour is removed by needle biopsy and is less invasive than SLNB [7]. A full factorial design is devised to consider different combinations of these markers. These combinations are then tested to find the most effective individual marker or combination set in nodal status prediction and to discover the existence of nonlinear inter-relations among these markers. Thus, this work offers the opportunity to investigate the ability of each feature set in differentiating node negative from node positive tumours using two different ANN structures.

Much research has been donated to the application of MLP structures in breast cancer prediction [17-20]. The reliability of resampling methods for error estimation in MLPs have also been investigated in some studies [28]. Nevertheless, no work has been done on the effectiveness of the PNN in breast cancer prediction using a reliable error estimation method. This research demonstrates that the complex classification problem of breast cancer prediction, which cannot be currently addressed by traditional research techniques, can be solved effectively by proper selection of biomarkers, the use of PNN as an apt choice of ANN architecture, and an appropriate selection of the PNN parameters and validation method. The problem chosen for this work is the prediction of ALN metastasis in breast cancer patients. The reason for this choice lies in the fact that this problem is one of the most complex problems in the field of pattern classification. The difficulty in carrying out research in this field is the large number of available tumour markers with a highly nonlinear relation to ALN status, in addition to the complex interrelations between these markers. This makes it difficult even for oncologists to predict nodal status from patient information. Moreover, the creation of an automated system would provide a huge benefit to the hospital field, which is constantly looking to reduce costs and mortality caused by unnecessary breast dissection.

The contributions of this study can be summarised as follow:

- The application of the PNN as a classifier for predicting the ALN metastasis and confirming the superiority of the PNN over the conventionally employed MLP through the comparison of both methods via a reliable error estimation technique
- Improving the reliability of error estimation for both the PNN and the MLP for breast cancer prediction using the .632 bootstrap technique
- accurately predicting the state of ALN metastasis in breast cancer and determining an effective set of biomarkers for nodal status prediction

One major drawback associated with the ANNs is that they are deemed as a black box which does not provide interpretable rules in the output. In recent years, many studies have focused on techniques to extract rules from ANNs [29, 30]. Nevertheless, this study focuses on addressing common deficiencies in previous ANN studies to obtain reliable results and a novel marker combination for predicting breast cancer metastasis to ALNs and the point is not toward extracting rules.

This study presents the results of two ANN structures evaluated with the .632 bootstrap estimator to determine a novel combination of tumour biomarkers for the prediction of ALN metastasis in breast cancer. The presented results for the prediction potential of the biomarkers in conjunction with the methods described in this research may serve as the basis for automatic and non-invasive prediction of breast cancer metastasis. This should significantly decrease patient morbidity by reducing the need for invasive procedures and decrease patient mortality by timely prediction of nodal metastasis.

1.6. Thesis outline

The overall organisation of this thesis is as follows: After the introduction, chapter 2 is dedicated to a review of the statistical, machine learning and ANN methods previously employed for breast cancer prediction. This chapter gives details of the advantages and disadvantages of the statistical and machine learning methods and the reason they have not been effective in breast cancer prediction. It then introduces the foundations of ANNs and a literature of the most widely used ANN structures in breast cancer diagnosis. The ANN outcome evaluation and currently employed error estimation methods are also described.

Chapter 3 introduces the biomarkers chosen in this study and the dataset characteristics, in addition to the biomarkers' measurement methods and descriptive statistics. The medical literature of the chosen biomarkers will be explained to establish their individual role and application in predicting the breast cancer progression. The final section of this chapter includes the scatter plots and correlation coefficients of the tumour biomarkers and the cancer metastasis to demonstrate the nature and the degree of correlation between different biomarkers and metastasis.

Chapter 4 starts with an explanation of the MLP structure, its training algorithm and common issues in cancer studies. Then an in depth explanation of the PNN, its structure and training algorithm is detailed. The last section of the chapter describes

.632 bootstrap which is an error estimation method based on resampling. The applied methods in conjunction with the employed data set bring about a novel application of PNN in breast cancer prediction.

Chapter 5 presents the experimental results of the study. In the first section of results, the method of .632 bootstrap error estimation is compared with the performance of holdout and 5-fold CV methods for both the PNN and the MLP. In the second section of results, the results achieved by the PNN and the MLP are applied to all marker combinations presented and are tabulated and analysed. Finally, the best results achieved by the PNN and the MLP are presented in a single table and compared against each other.

Chapter 6 contains a discussion of the results including an analysis of the reliability of different error estimation methods, a comparison of two ANNs and an analysis of results in terms of biomarkers' prediction potential. The last part of discussion is dedicated to explain how the disadvantages in previous studies have been addressed in this work. Finally the conclusions drawn from this study, the contributions of the research and possible future research directions in this field are detailed in chapter 7.

Chapter 2

2. Statistical and classification methods for breast cancer prediction

2.1. Introduction

Clinical decision making is based on multiple symptoms and measurements obtained from the patient. Physicians make diagnosis considering these multiple factors acquired from the patient, assigning weights to each factor and choosing the most probable diagnosis. This is usually done unconsciously by the physician based on the relative importance of each factor [31]. Physicians typically obtain this knowledge through experience by examining several patients, making diagnosis and then comparing it with the actual outcome. Therefore, experienced physicians are more likely to make a correct diagnosis than the novices.

However, there are situations that even an expert needs assistance to make the correct decision. This is especially true for cancer prediction as there is excessively large number of factors discovered to be related to the outcome. In addition, the relation between many of these factors and the cancer outcome in terms of metastasis or survival is not completely revealed yet. Availability of experts is another issue which necessitates the use of decision support systems (DSS) for cancer prediction.

DSSs offer numerous advantages in medical decision making by taking into account a large number of factors simultaneously such that the complex relationship between the factors and the cancer outcome can be distinguished. This brings a significant benefit to medical centres in terms of costs and availability of expert oncologists.

There are different types of medical DSSs depending on their approach to data classification. The main methodologies employed in breast cancer prediction are statistical and machine learning methods [32] for which a literature survey is reported in this chapter. For this purpose, regression and discriminant analysis (DA) are explained as the most commonly applied statistical methods and decision tree (DT) as the most routinely employed machine learning method in breast cancer prediction. In addition, the advantages and disadvantages of these methods are detailed. This leads the discussion to the necessity of employing more sophisticated classification algorithms such as ANN for overcoming the shortcomings of statistical and machine learning methods such as assumptions about underlying distribution of data and requiring large sample size in each output class.

After the literature survey on the statistical and machine learning methods employed in breast cancer prediction, the foundations of ANNs, how they are trained and their learning process are explained. Then, a comprehensive literature of ANN application in breast cancer prediction, its advantages, disadvantages and current problems are mentioned and analysed. This elucidates the main gaps in the applied ANN systems to breast cancer prediction which forms the focus of the thesis and the topics to be tackled in the next chapters.

2.2. Statistical methods in cancer prediction

Conventional statistical methods used in breast cancer prediction are regression methods and DA. These methods are mainly employed for classification of patients into different outcome groups using a set of discriminatory factors. A history of these techniques, their principles, advantages and disadvantages are detailed below.

2.2.1. Regression Models

Regression is the most commonly used statistical method in medical studies. It is employed to model and assess the relationship between a single or a set of input and output variables. In regression analysis, inputs are described as predictors or independent variables, while the outputs are termed as predicted outcome or dependant variables.

There are different regression models depending on the type and number of input and output variables and their relationship. Linear regression, models the input-output relation with a straight line while in a curvilinear regression model, any function forming a curve line can be used for input-output modelling. Multivariate regression analysis can be employed in case of developing a model for multiple inputs with a single output.

After developing the regression model for a set of input and output variables, the model is assessed to study how well the model fits the data and whether it requires any modifications. In addition to predictive modelling, regression analysis can be utilised as an exploratory tool to understand the nature of relationship between multiple factors. This helps to verify any existing relationship between variables and to clarify the description of this relationship.

Linear regression is the simplest form of regression analysis. As mentioned before, it describes the linear relation between a dependant and independent variables. Prior to linear regression, it is common to use another measure of linear relation, called correlation coefficient, to measure the linear dependency of two variables. Correlation coefficient does not consider any causal relationship between the variables and it merely quantifies the strength of the linear relation between them.

In the following sections, correlation coefficient is explained followed by linear regression analysis. LR is then explained in detail including its model building, underlying assumptions and the interpretation of the fitted model.

2.2.2. Correlation coefficient

Correlation coefficient is a measure of linear dependency between two variables. It is commonly employed prior to linear regression to explore the existence of any linear relationship between the dependant and independent variables. The most commonly used method for quantifying the amount of linear dependency between two continuous variables, A and B, is Pearson’s Correlation Coefficient (PCC). PCC

represents the strength of linear association between two variables A and B by normalising their covariance with respect to their standard deviation S_A and S_B as:

$$PCC_{A,B} = \frac{cov(A,B)}{S_A S_B} = \frac{E\langle(A-\mu_A)(B-\mu_B)\rangle}{S_A S_B} \qquad (2.1)$$

where μ_A and μ_B are the expected values of two random variables A and B. PCC assigns a number between -1 and 1 for the measure of linear dependency between variables. A positive value represents a positive linear relationship, while a negative one implies negative linear and 0 suggests no relationship between variables.

The main characteristic of the PCC is its invariance to linear transformation of the variables A and B. This implies that for example, with variables being described as $A = 5B + 3$, PCC between the two variables represent a perfect correlation of *1*.

2.2.3. Linear regression

Linear regression describes the relationship between an independent variable X and a dependant variable T given a set of data points(x_i, t_i). This relationship is expressed as a straight line of the form:

$$t_i = \beta_0 + \beta_1 x_i + e_i \qquad (2.2)$$

where β_0 and β_1 are the regression coefficients describing the intersect and the line slope respectively. The term e_i is the error defined as the difference between the data point (x_i, t_i) and the regression line. Linear regression analysis is performed to estimate the unknown coefficients β_0 and β_1. Least squares method is commonly used for estimating coefficients to fit a regression line to the data. In this method, the aim is to minimise the sum of squared errors, i.e. finding the solution to the derivative of $\sum e_i^2$:

$$\sum e_i^2 = \sum_{i=1}^{N} [t_i - (\beta_0 + \beta_1 x_i)]^2 \tag{2.3}$$

Getting the derivative of $\sum e_i^2$ with respect to β_0 and β_1 leads to the following equations:

$$\frac{\partial \sum e_i^2}{\partial \beta_0} = -2 \sum_{i=1}^{N} [t_i - (\beta_0 + \beta_1 x_i)] = 0 \tag{2.4}$$

$$\frac{\partial \sum e_i^2}{\partial \beta_1} = -2 \sum_{i=1}^{N} x_i [t_i - (\beta_0 + \beta_1 x_i)] = 0 \tag{2.5}$$

Solving the above equations with respect to the two unknown variables β_0 and β_1 gives [33]:

$$\beta_0 = \frac{\sum_{i=1}^{N} t_i \sum_{i=1}^{N} {x_i}^2 - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} x_i t_i}{N \sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{\bar{t}\left(\sum_{i=1}^{N} {x_i}^2\right) - \bar{x} \sum_{i=1}^{N} x_i t_i}{\sum_{i=1}^{N} {x_i}^2 - N\bar{x}^2} \tag{2.6}$$

$$\beta_1 = \frac{N \sum_{i=1}^{N} x_i t_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} t_i}{N \sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = \frac{\left(\sum_{i=1}^{N} x_i t_i\right) - N\bar{x}\bar{t}}{\sum_{i=1}^{N} {x_i}^2 - N\bar{x}^2} \tag{2.7}$$

The covariance and standard deviation (S_x) definitions are defined as:

$$cov(x, t) = \sum_{i=1}^{N} (x_i - \bar{x})(t_i - \bar{t}) = \sum_{i=1}^{N} x_i t_i - N\bar{x}\bar{t} \tag{2.8}$$

$$S_x = \sum_{i=1}^{N} (x_i - \bar{x})^2 = \sum_{i=1}^{N} {x_i}^2 - N\bar{x}^2 \quad (2.9)$$

where $\bar{x}$ and $\bar{t}$ represent the mean of the vectors x and t respectively. Using the covariance and standard deviation, $\beta_1$can be rewritten as:

$$\beta_1 = \frac{cov(x,t)}{S_x} \quad (2.10)$$

Having β_1, β_0 is defined from (2.4) as:

$$\beta_0 = \bar{t} - \beta_1\bar{x} \quad (2.11)$$

Quality of fitting the line to the data is then quantified and assessed by the correlation coefficient as defined in equation (2.1). Least square is the most common method to estimate the regression coefficients as it is equivalent to maximum likelihood estimation under the assumption of normally distributed error. Multivariate linear regression is an extension of simple linear regression having more than one independent variable. The relation between the dependant and independent variables is described in linear form as:

$$T = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_P X_P + e \quad (2.12)$$

where β_0 is a constant and β_1 to β_P are the coefficients of P independent variables.

Linear regression is limited to continuous dependant variables. This limitation prohibits its application in classification problems with dichotomous outputs such as medical studies with a binary output of the presence and absence of the disease. LR is the most common statistical technique utilised to overcome this limitation.

2.2.4. Logistic regression

To overcome the limitation of linear regression in handling dichotomous classification problems, LR was proposed in early 1970s and become widely available to researchers through statistical software in early 1980s [34].

In medical diagnosis, where the disease outcome is described as present or absent, the outcome is a binary variable taking values of 1 or 0 similar to the LR classification outcome. Hence, LR has been routinely applied and became a standard method in the field of medical diagnosis as a technique which is simple to apply through several statistical packages. In breast cancer, LR has been frequently used to predict dichotomous outcomes of patient survival, diagnosis and prognosis [35-38].

As a linear multivariate analysis technique, LR describes the relation between one or more independent variables and a binary outcome. LR employs the same principles as linear regression for data analysis. Like linear regression, a model is developed to fit the input-output data by estimating a set of coefficients. However, because the output can take any value from a continuous range of outcomes in linear regression, it is difficult to assign a dichotomous output to (2.2). For this reason, logistic distribution is used for describing the input-output relationship that is defined as the conditional mean of the output T given the input x [39]:

$$\pi(x) = E(T|x) = \frac{e^{\beta_0+\beta_1 x}}{1 + e^{\beta_0+\beta_1 x}} \tag{2.13}$$

Since the output is dichotomous (i.e. $T = 0,1$), $\pi(x)$ can be described as the probability of the output happening $(T = 1)$. Similarly, the probability of the output not happening $(T = 0)$ is defined as $1 - \pi(x)$. For final results, the output in (2.13) is transformed using a *logit* transformation. The logit transformation is the inverse of logistic function and is defined as the natural logarithm (ln) of the odds of the output. The odds are the ratios of $\pi(x)$ to $1 - \pi(x)$. The LR output $g(x)$ is therefore defined as:

$$g(x) = \ln\left(\frac{\pi(x)}{1 - \pi(x)}\right) = \beta_0 + \beta_1 x \qquad (2.14)$$

Hence, instead of directly predicting the output from input variables, the logit of the output is predicted from the input. The equations (2.13) and (2.14) confirm that while the relation between the probability of the output, $\pi(x)$ and the input x is nonlinear; the relation between the logit of the output T and the input x is a linear one.

Simple LR can be extended to multivariate LR by expanding the input x to a vector of multiple input factors $X = \{X_1, X_2, ..., X_p\}$:

$$\pi(x) = \frac{e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_P X_P}}{1 + e^{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_P X_P}} \qquad (2.15)$$

The coefficients in LR are commonly estimated by maximum likelihood. In this method, the probability of the observed data as a function of the unknown coefficients is defined as a likelihood function. The unknown coefficients are chosen such that the likelihood function is maximised [39].

LR regression outcome $g(x)$ retains some characteristics of linear regression such as linearity in parameters and continuity within the range of $-\infty$ to $+\infty$ depending on the range of x. Nevertheless, in LR conditional distribution of the outcome follows a binomial distribution [39]. One of the important properties that make LR the desirable method of choice in medical studies is that it can be easily implemented and it also provides meaningful interpretation of the results.

LR makes no assumptions about the underlying distribution of the independent variables. However, LR only models linear relationships between the dependant and independent variables. Besides, for a good estimate of the LR parameters, the independent variables should not be highly correlated. In addition, a fairly large sample size is required to supply sufficient data in each of the output classes for LR analysis. Sample sizes greater than 400 are suggested for each treatment group to build a reliable LR model [39].

2.2.5. Discriminant Analysis

The purpose of DA is to predict the class membership of a data point using a set of predictive variables. Simplest form of DA, linear discriminant analysis (LDA), was developed in 1936 by Fisher for data analysis and classification [40]. Similar to LR, LDA is suitable to be applied for linear classification problems with dichotomous outputs. For this reason, it can be a suitable method in medical diagnosis for identifying differences between patients with and without ALN metastasis and classifying them into one of these groups.

LDA classifies data through maximising data separablility by projecting it into a new space where groups are maximally separated. Fisher LDA finds the transformation vector w such that the distances between the means of transformed inputs are maximised while the variances of the projected data in each group are minimised.

The means and variances of the original and the transformed data are defined as below:

$$\mu_i = \frac{1}{N_c}\sum_{i=1}^{N_c} x_i \qquad (2.16)$$

$$\widetilde{\mu_\iota} = \frac{1}{N_c}\sum_{i=1}^{N_c} w^T x_i = w^T \mu_i \qquad (2.17)$$

$$s_i{}^2 = \sum_{i=1}^{N_c}(x_i - \mu_i)(x_i - \mu_i)^T \qquad (2.18)$$

$$\widetilde{s_\iota}{}^2 = \sum_{i=1}^{N_c}(w^T x_i - \widetilde{\mu_\iota})^2 = \sum_{i=1}^{N_c}(w^T x_i - w^T \mu_i)^2 = \qquad (2.19)$$

$$\sum_{i=1}^{N_t} \mathrm{w}^T (x_i - \mu_i)(x_i - \mu_i)^T w = w^T s_i w$$

where $\widetilde{\mu_1}$ and $\widetilde{\mu_2}$ are the means of transferred data in each group and $\widetilde{s_1}$ and $\widetilde{s_2}$ are their standard deviations (~ shows that the values are associated with the transferred data). In order to find a transformation vector w such that the distances between the means of transformed inputs are maximised while the variances of the projected data in each group are minimised, the function to be maximised for a two class problem is formulated as:

$$C(w) = \frac{|\widetilde{\mu_1} - \widetilde{\mu_2}|^2}{\widetilde{s_1}^2 + \widetilde{s_2}^2} \qquad (2.20)$$

The obtained transformation vector w is then applied to the input vector x to transform it into a new space $w^T x$.

In order to obtain w from (2.20), the derivative of $C(w)$ with respect to w is computed. For this reason, $C(w)$ must be expressed in terms of w and the original data. Hence, the *between-class* and *within-class* scatter matrices, S_B and S_W are introduced as below:

$$s_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T \tag{2.21}$$

$$S_W = S_1 + S_2 \tag{2.22}$$

And therefore, the numerator and denominator of (2.20) are presented as:

$$(\widetilde{\mu_1} - \widetilde{\mu_2})^2 = (w^T\mu_1 - w^T\mu_2)^2 = w^T(\mu_1 - \mu_2)(\mu_1 - \mu_2)^T w = w^T S_B w \tag{2.23}$$

$$\widetilde{s_1}^2 + \widetilde{s_2}^2 = w^T S_W w \tag{2.24}$$

$$C(w) = \frac{|\widetilde{\mu_1} - \widetilde{\mu_2}|^2}{\widetilde{s_1}^2 + \widetilde{s_2}^2} = \frac{w^T S_B w}{w^T S_W w} \tag{2.25}$$

The transformation w is then obtained by solving the derivation of $C(w)$ with respect to w. Afterwards, a new datum can be classified by being transformed by w and measuring the distance between the transformed datum and the mean of the transformed data in each group. The new datum is then assigned to the group which obtains the minimum distance.

The main limitation of LDA in medical applications is that it only deals with continuous input variables and makes more assumptions about the input data compared to other statistical methods such as LR. LDA assumptions about independent variables are multivariate normality and equal covariance. If these conditions are satisfied, LDA gives better or similar results compared to LR [41]. However, medical datasets are unlikely to lend themselves to these properties. Furthermore, it has been proved that LR performs better than or similar to LDA in most applications [41]. Therefore, the application of LDA in medical studies is limited.

Some variations of DA have been proposed to obviate the limitations of LDA such as nonparametric LDA. In this method, the limitation of Gaussian assumption is removed by employing the k-nearest neighbours rule for computing the between-class scatter matrix [42]. However, this method still has limitations when the between-class scatter matrix is singular.

2.3. Problems with statistical methods

Using more than one predictor is common in cancer studies as it is unlikely that one factor can provide discrimination between patients with and without a specific outcome. For this reason, multiple predictors are employed to classify the cancer outcome which gives rise to a multidimensional input space. The relationship between multiple predictors is typically a nonlinear and complex one which renders the linear statistical methods insufficient for these classification problems. In additions, many statistical methods make prior assumptions about the data such as multidimensional normality that is not generally true in cancer studies.

An example of a nonlinear classification problem with two predictors (i.e. 2-dimensional input space) and a binary output variable is shown in Figure 2.1 (a) and (b). In these figures, problems with nonlinear decision boundary and closed curve boundary are presented [43]. In both cases, a linear classifier such as LR or LDA would fail to separate the data. In order to employ linear statistical methods for problems with nonlinear decision boundaries, data can be transformed nonlinearly into a new space where it is more separable and methods such as LR or LDA can be used. Then again, there is no guideline on how to choose an optimal nonlinear transformation and hence, it is unlikely to find a transformation that projects data into a space where it is completely separable.

ANNs can solve this problem by automatically determining the data transformation and finding an optimal decision boundary. MLP is a specific ANN architecture which resembles a form of nonlinear DA that can overcome LDA limitations by transforming the inputs in a nonlinear fashion. Like DA methods, ANNs utilise training data with known outcome to build a model with which a new datum can be classified. However, they make no or little assumptions about the data and are able to learn nonlinear patterns from training data [44]. For this reason, they are chosen as a desirable classification technique in many real-world problems [45, 46]. In addition, ANNs yield higher accuracy in cancer diagnosis and prognosis compared to conventional statistical methods such as LR and Bayesian classifier [13, 14, 47, 48]. ANNs are the main platform used in this study and are fully discussed in the later sections of this chapter.

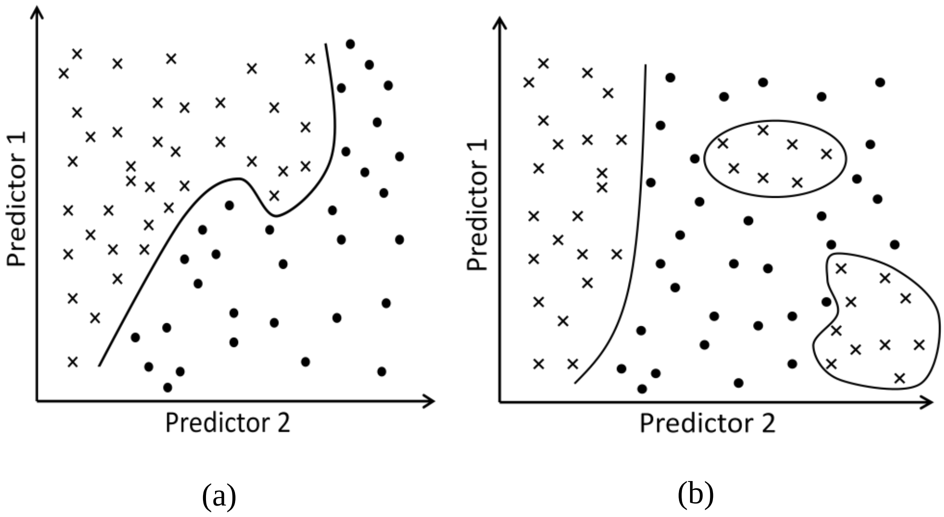

Figure 2.1: Nonlinear classification problem with two predictors and a binary output with (a) nonlinear decision boundary and (b) closed curve boundary. In both cases, linear classifiers fail to separate the data.

2.4. Machine learning techniques in cancer diagnosis

Machine learning is a branch of artificial intelligence that embraces a wide range of methods including DT, rule induction methods, Bayesian learning, genetic algorithms and ANNs [49]. These methods employ various statistical and optimisation techniques to provide decision making by learning complex patterns in a population from a sample set of data. Because of this characteristic, machine learning techniques have been widely applied in the field of medical diagnosis and prognosis.

Machine learning methods generally achieve 15-20% improvement in accuracy for predicting cancer survival and recurrence besides providing more insight about the nature of cancer progression. Among various machine learning techniques, DTs are among the most widely used methods in cancer diagnosis [50]. DTs are successfully applied to ovarian and gastric cancers [51, 52]. However, DTs are usually used in conjunction with another classification method to achieve better classification results [53, 54]. A brief history of DT in medical literature, its structure and learning algorithm, in addition to its advantages and disadvantages are explained in the following section.

2.4.1. Decision tree (DT)

DT learning is an inductive inference algorithm that can approximate discrete outcomes using a set of training data. The learned function by DT is represented by a set of branches and nodes that are structured as a graph. The graph starts from a root node that is attributed to one of the input variables followed by leaf nodes attributed to each of the remaining input variables. The outcome of the tree is a category into which the input is classified. An example of the structure of a DT for a simple classification problem is presented in Figure 2.2 [49]. In this example, the aim is to decide whether to play tennis or not. The decision depends upon the weather forecast and is 1 if:

Outlook = Overcast

OR Outlook = Sunny AND Humidity = High

OR Outlook = Rain AND Wind = Weak

The output would be otherwise *0*.

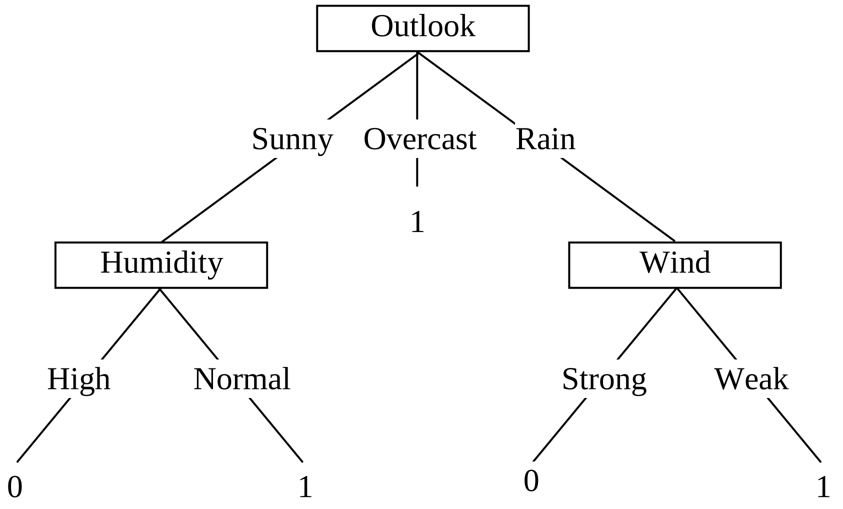

Figure 2.2: A decision tree structure to decide whether to play tennis or not based on weather forecast. The output is either *1* for play tennis or *0* for not play tennis.

The simplest form of learning in DT consists of a greedy search that starts from the root node and only goes down without backtracking to the previous nodes [49]. The root node is selected using a predefined measure. The training data are then assigned to the succeeding leaf nodes based on their values of the root node variable. The leaf nodes' variables are then selected using the same measure and a similar process is repeated until a decision is reached or no further division is possible. In ID3 decision trees, the measure for choosing the best attribute in each step is *information gain* [55]. This measure quantifies how well each variable classifies training data into output categories. Information gain is defined using a well known measure in information theory called *entropy*. The entropy of the dataset *X* in respect to a two category output class is defined as:

$$Entropy(X) = \sum_{c=1,2} -P_c log_2 P_c \tag{2.26}$$

where P_c is the proportion of the data points in X belonging to class c. Entropy can take values between 0 to 1 depending on the proportion of each class in the data. If all the data belongs only to one category, the entropy is equal to 0 while it is 1 when there are equal numbers of each category. An unequal number of output categories result in an entropy value between 0 to 1. The information gain of a variable is then defined as:

$$IG(X,V) = Entropy(X) - \sum_{v \in Values(V)} \frac{|X_v|}{|X|} Entropy(X_v) \tag{2.27}$$

where V is the input variable and $Values(V)$ are its possible values. X_v is the subset of X for which variable V has values v.

DTs maintain several advantages that have made them attractive in medical diagnosis. The results provided by DTs are easy to interpret in terms of if-then rules. In addition, DTs are considered as robust classifiers with quick learning [49]. However, DTs do not perform as well as ANNs in complex classification problems [56].

One limitation in DTs is that each attribute is required to have a small number of discrete values. To address this problem, continuous variables are partitioned into discrete intervals. A new dummy variable is then introduced that takes discrete values attributed to each of the intervals. The main disadvantage of implementing continuous inputs in DT is the loss of information that occurs due to partitioning.

Another problem that arises with partitioning the continuous variables is selecting the appropriate intervals.

Furthermore, DTs are highly prone to overfitting. DTs may present good predictions for small training data with missing values. In effect, a large tree may separate the training data perfectly with no misclassification error. However, such model may overfit the data by having few training samples at the end of each branch which means low generalisation ability for new data.

Support vector machines (SVMs) are another machine learning technique which can be classified as Kernel-based methods [55]. SVMs are supervised learning methods based on the statistical learning theory and the Vapnik-Chervonenkis dimension [56]. For a linear classification problem with dichotomous outcomes, SVM finds the decision boundary in the form of a hyperplane such that it is as far as possible from the closest members of both classes. In nonlinear SVMs, the data are recast into a higher dimensional space by using a nonlinear kernel function such that they are linearly separable in the new space. It has been shown that SVM performs as good as ANNs such as the PNN in breast cancer detection [57]. However, using a nonlinear SVM means that an appropriate kernel function and its parameters should be chosen such that the nonlinearly separable data can be mapped into a higher dimensional space where it is linearly separable. Unfortunately, there are no specific rules for choosing such function and it is commonly chosen using trial and error from a set of predefined kernel functions. This makes the success of SVMs dependant on finding an appropriate transfer function which might not be possible for highly nonlinear and complex data.

2.5. Neural Network Models

DSSs for cancer diagnosis have been widely employed in clinical practice. A considerable number of these systems involve ANNs. A wider context of DSSs in medical diagnosis and other methods routinely employed in cancer diagnosis including statistical and machine learning methods have been already covered in the previous sections. In this section, a specific review of ANNs and their application in cancer diagnosis is detailed.

ANNs' basic form and the preceptron concept were first introduced by McCulloch and Pitts in early 1940s as a tool for simulating the human intelligence [57]. ANNs exploit special characteristics of human brain such as reasoning, making decision and learning by experience. ANNs are proved as valuable classification tools in many applications such as business, science and engineering [46]. They maintain several advantages over conventional statistical methods especially for data that exhibits nonlinearity. The vast amount of research conducted in this field has established that ANNs perform better than various statistical, machine learning and rule-based methods such as LR, DA and DT [58-60]. This is due to the special theoretical properties of ANNs including:

- ANNs are nonlinear models. Hence, they are suitable for real world applications and especially cancer prognosis in which there exists a nonlinear and complex relationship between the biomarkers and cancer outcome. This nonlinear modelling is established in the following sections where the theoretical principles of ANNs are explained.

- ANNs can be trained to learn data patterns by means of adaptable weights. This makes them suitable for data driven analysis without making any prior assumptions about the underlying models.

Because of these advantages, ANN systems have been employed as medical DSSs from the time when early computing systems where introduced [31]. Since then, there has been wide research and many improvements on the application of ANN in medical diagnosis. However, it has been only in the past two decades that ANNs have been practically applied in the field of medical decision making [36]. This has helped many physicians in making medical decisions more confidently and in a timely fashion.

2.5.1. An overview of artificial neural networks

ANNs are a branch of artificial intelligence that are able to learn complicated nonlinear patterns from a set of data. ANNs are parallel computational units which have become a valuable classification tool in recent years. They first originated from the idea of simulating human brain abilities in decision making and parallel processing by combining mathematical modelling and engineering design. While the basic structure and characteristics of ANNs still resemble the human brain, their functioning and the way they make decisions have become far different from biological neural networks over time.

Different types of ANNs can be categorized by two main criteria [61]. First criterion is how the network is encoded, i.e. how the network stores knowledge from the data. Using this measure, ANNs are categorised as supervised and unsupervised. Second criterion is the way the networks are decoded, i.e. the way the network processes new data once it has acquired knowledge from the old data. This criterion classifies ANNs into feedforward and feedback. Figure 2.3 illustrates this classification.

In a supervised network, both the input and output are presented to the network. The network weights are then adjusted by computing an error from comparing the network output and the desired output. The optimum weights are obtained by optimising the error function. Afterwards, the ability of the network in classifying new data is tested by presenting new inputs to the network and comparing the answer with the unseen output. Some ANNs have the ability to learn without teachers. The learning process when only inputs are presented to the network is termed as unsupervised learning. This is achieved by using rules for self-adjustment as the new inputs are presented to the network. The self-adjustment is performed based on predefined rules [62].

		Decoding	
		Feedforward	Feedback
Encoding	Supervised	Radial-Basis Function (RBF) Networks Multilayer Perceptron (MLP)	Boltzmann Machine (BM)
	Unsupervised	Self-Organising Maps (SOM)	Hopfield Networks

Figure 2.3: Different types of ANNs can be categorized by two main criteria: how they are encoded, (supervised and unsupervised) and how they are decoded (feedforward and feedback).

Excessive and continuously growing literature is available on ANNs [46, 62-65]. The theory of ANNs is derived from different disciplines including mathematics, statistics, biology, engineering, computer science and neuroscience. Each of these disciplines contributes to the capability of ANNs as intelligent systems to be employed in a wide range of applications. A comprehensive explanation of the mentioned disciplines forming the foundations of ANNs is out of the scope of this thesis. Therefore, the focus is only on the paradigms that contribute to the ANNs applied in the field of medical diagnosis and more specifically, to breast cancer prediction. In the following sections, the fundamentals of the ANNs are briefly explained to prepare the grounds for further explicating the ANN models used in this project.

2.5.2. Neurons

As mentioned above, the structure and function of ANNs resemble the human brain, in that they include a massively parallel architecture, but in a relatively small scale. Whilst human brain is made up of 10^{11} neurons in average, ANN is based on layers of computing nodes which include only a few computing units.

Biological NNs consist of controlling units called neurons which have the ability to learn and work in parallel. ANNs emerged in 1943 for the first time by representing a simple form of biological neurons by elements that could perform computation [57].

Biological neurons have different specialisation and functioning. A simplified structure of a Biological neuron consists of four main parts. Each neuron has a cell body called soma which can receive input from nearly 10000 other units through an assembly of subtle structures called dendrites. After processing the received signal, the neuron sends out an electrical potential through a long, thin structure called axon which divides into many branches. Afterwards, the electrical output is transferred to the dendrites of other neurons via junctions called synapse. This structure is illustrated in Figure 2.4.

.

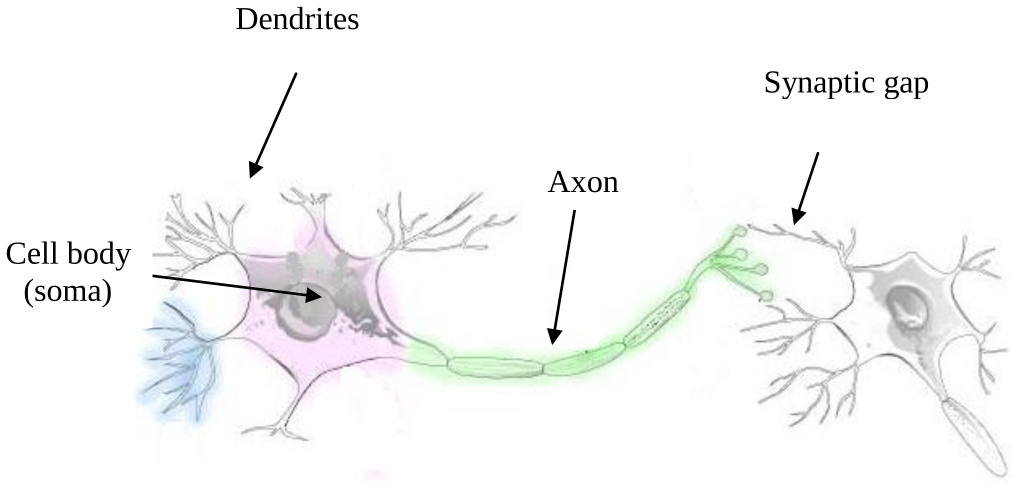

Figure 2.4: A biological neuron

For creating an ANN model based on a biological NN, there are three important components of a neuron to simulate. The synapses are represented as weights which define the strengths of the connection between nodes. The dendrites and axon are modelled as the actual activity taking place in ANN nodes where all the inputs are summed up by considering their weight through a process called linear combination. Finally, the output amplitude of the neuron is defined by an activation function.

An artificial neuron can be considered as a simplified biological neuron in that it has the ability to learn from previous data and predict the output for new data. However, the processing rate, size, complexity level and the fault tolerance of them is not comparable. A simplified model of an artificial neuron is illustrated in Figure 2.5 in which n inputs denoted by X_i are weighted through w_i weights and summed up by Σ, then passed through an activation function indicated by φ and transferred to the neuron output expressed by O. An artificial neuron also contains bias which is an external input for adjusting the net input of the activation function. Bias can be considered as the 0th input weighted by a unit connection. Hence, the output of a neuron can be formulated as:

$$O = \varphi(\sum_{i=0}^{N} w_i \, x_i) \tag{2.28}$$

Where the sum of the weighted inputs ($\sum_{i=0}^{N} w_i \, x_i$) is called the *activation* of the neuron.

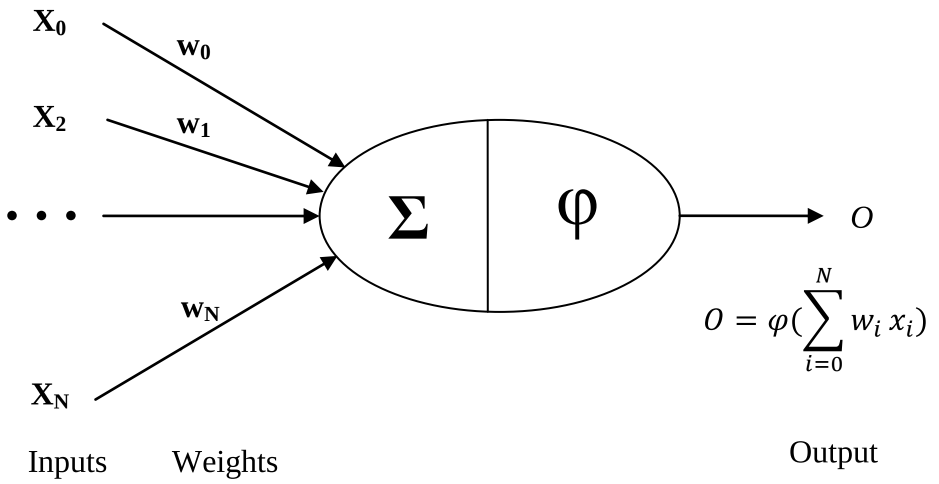

Figure 2.5: A simplified model of an artificial neuron

Activation function defines the output level of the neuron for a given input. There are three general types of activation functions namely identity, threshold and sigmoid transfer functions. When the net input weights to the neuron are transferred to the output directly, without any changes, the neuron is considered to have an identity or linear transfer function. A linear activation function can be simply formulated as:

$$\varphi(v) = v \tag{2.29}$$

with v being the activation of a neuron. In threshold activation functions, the output takes the value of $+1$ if the input is larger than a defined threshold t and takes the value of 0 otherwise:

$$\varphi(v) = \begin{cases} 1 & if v > t \\ 0 & if v \leq t \end{cases} \tag{2.30}$$

Some activation functions such as logistic or hyperbolic tangent are considered as sigmoid transfer functions. These functions have bounded range in the output; hence, they are also referred as ‘squashing’ functions. Logistic function is defined by:

$$\varphi(v) = logsig(\mathrm{s}(v + b)) = \frac{1}{1 + e^{-\mathrm{s}(v+b)}} \tag{2.31}$$

where *s* is the steepness parameter defining the slope of the function and *b* is the bias of the function defining its location on the horizontal axis.

Similarly, a hyperbolic tangent function is defined by:

$$\varphi(v) = tanh(\mathrm{s}(v + b)) = \frac{e^{2\mathrm{s}(v+b)} - 1}{e^{2\mathrm{s}(v+b)} + 1} \tag{2.32}$$

These transfer functions are illustrated in Figure 2.6.

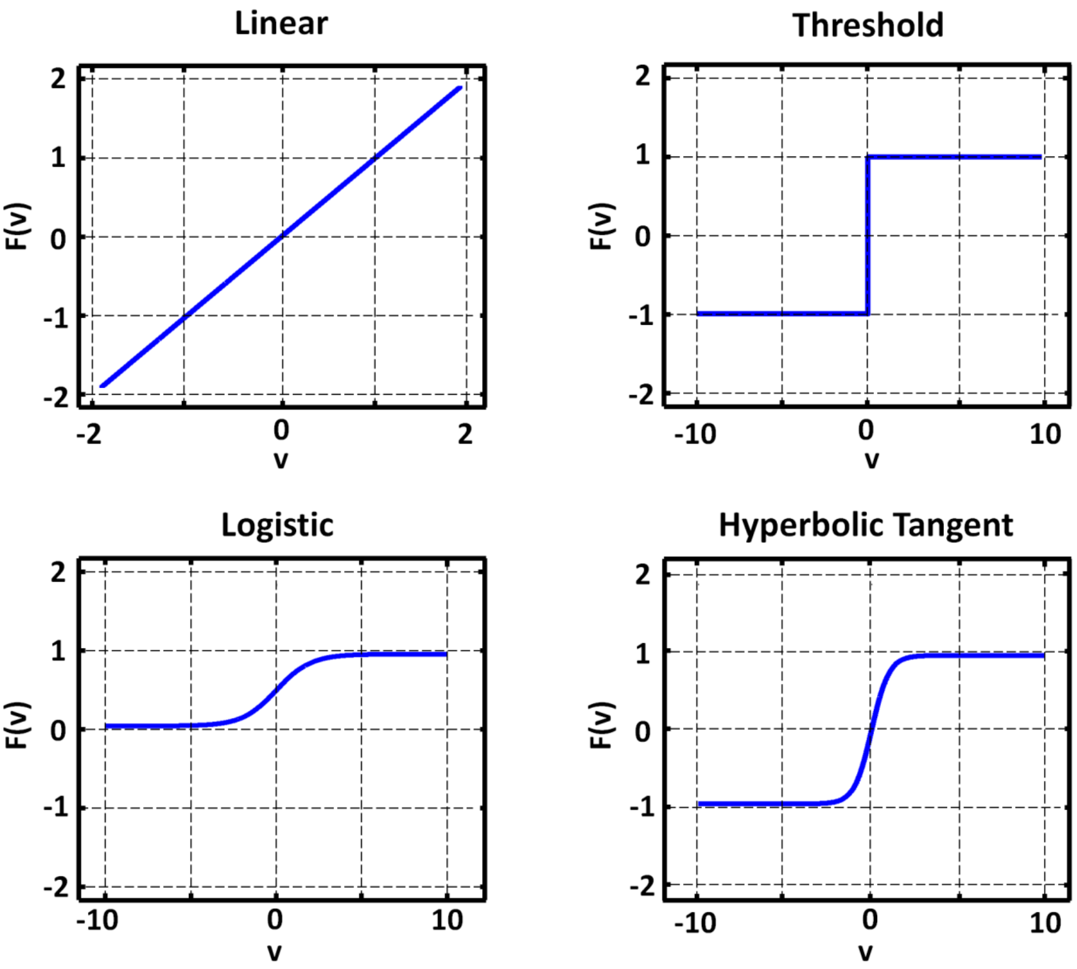

Figure 2.6: Linear, threshold, logistic and hyperbolic tangent transfer functions

Sigmoid transfer functions are commonly used in ANNs because of their special mathematical properties. These properties include continuity, differentiability at all points and monotonicity (i.e. monotonically increasing in a finite range). Among Sigmoid functions, hyperbolic tangent is preferred over logistic function because of its symmetrical output.

2.5.3. Rosenblatt’s perceptron

The first model for supervised learning called perceptron was introduced by Rosenblatt in 1958 [66]. Rosenblatt’s perceptron is still valid as the simplest form of an ANN for classifying linearly separable patterns (i.e. patterns that can be divided into two distinct classes via a hyperplane).

Perceptron consists of a single neuron with adaptable weights and bias. Training in perceptron is inspired by biological neural systems. This is accomplished by reinforcing good behaviour (correct output) and discouraging bad behaviour (incorrect output). In perceptron, reinforcing and discouraging is simulated by adapting weights of the connections linking the inputs to the neuron. The amount of weight change is decided by considering the distinction between the network and desired outputs.

In order to measure the degree of closeness of the network output and the desired output, network’s error at step s is computed by subtracting the network’s output *o* from the desired output *t*:

$$e(s) = t(s) - o(s) \qquad (2.33)$$

The weights are then updated as follow:

$$w_i(s+1) = w_i(s) + \beta \mathrm{e}(t) \mathrm{x_i}(s) \qquad (2.34)$$

In the above equation, β is called the learning rate which has the effect of controlling the rate of change in the weights. This means taking smaller steps when the trained network is close to the desired solution and taking larger steps otherwise.

The process of weight adaptation is started by initialising the network weights randomly to a small value. The network output is computed accordingly and error is computed from (2.33). Then the weights are updated from (2.34) and the same process is repeated for several steps where the error is expected to reduce after each step.

Rosenblatt proved that given two linearly separable classes, the perceptron algorithm converges to a hyperplane separating the input patterns into two distinct groups in a finite number of steps. This is known as the perceptron convergence theorem. A detailed proof of perceptron convergence theorem can be found in Haykin [62].

The perceptron algorithm is found to be insufficient in some aspects. One issue is the myriad number of solutions existing for a single linear problem depending on the initial weights. The second issue is that the number of steps needed to reach the optimal solution can be very large. Finally, when the problem at hand in nonlinear, the algorithm does not converge and remains in an indefinite loop of steps. To overcome these limitations, MLPs are introduced and structured as fully connected layers composed of perceptrons.

Another type of feed-forward ANNs are radial basis function (RBFs). RBF neural networks are the main alternative to the MLP for nonlinear classification. A special design of the RBF for dichotomous outputs, called PNN, is devised as the main platform employed in this study. The MLP and the PNN, their structure, learning algorithm, advantages and disadvantages are detailed in chapter 4. These ANNs constitute the main approaches in this thesis and therefore, a history of their application in cancer prediction and diagnosis is entailed in the next section.

2.6. Neural network models for breast cancer prediction

There has been significant number of publications on ANN approaches applied as DSS in cancer diagnosis and prognosis in the last two decades. An introduction to ANN applications in oncology and the impact of ANNs on cancer research can be found in [67]. A Lancet series published in 1995 covered the principles of ANNs and their application in classifying clinical data [32, 43, 68]. Numerous studies have confirmed the efficiency and reliability of ANNs in cancer prognosis [69]. Burke et. al. [15] has employed a MLP trained with BP and compared it to the conventional cancer prognosis method for 5-year survival prediction and has demonstrated the significant superiority of ANN.

In a comparison between regression models and ANNs applied to classifying medical data, it has been suggested that with small datasets (less than 2000), ANNs outperform regression models [59]. In another comparison study, Burke et al. [14] advocate the superiority of ANN to linear and logistic regression and demonstrate that the performance of the MLP and the PNN in cancer prognosis are almost identical. Main reason for the better performance of ANN in the mentioned studies is the nonlinear and complex nature of input-output relation in cancer datasets. Hence, the classification analysis in cancer prediction is a nonlinear one which necessitates having efficient nonlinear classifiers. This is the primary reason for using ANN as the main predictive model in this study instead of conventional statistical tools.

For the specific application of ANNs in prediction of ALN metastasis in breast cancer, a comprehensive review has been conducted by Patani et al. [7]. They have surveyed the studies investigating the prediction potential of biomarkers for determining the ALN status. This review only entails those articles meeting five criteria including a clarified and complete explanation of tumour characteristics, study design, methods, results and the test significance (with P-value<0.05). Considering these criteria, only 30 articles out of 290 were included in the review which reveals the large number of unapproved studies in this field. Among these 30 articles, only two were conducted using ANN. Both publications had employed MLP neural network for the analysis of a set of biomarkers to predict the ALN status [70, 71]. It has been concluded that ANN based analysis of multiple biomarkers is advantages over conventional analysis techniques such as LR [7]. Other studies have also confirmed the capability of ANNs in providing accurate and reliable prediction for ALN metastasis status utilising tumour biomarkers [72].

Among various ANN architectures, MLP has been the most widely used method so far for cancer prediction and prognosis [20-22]. The effectiveness of MLPs in breast cancer diagnosis has been evaluated using clinical, pathological and immunohistochemical data which suggest MLP as a powerful technique for cancer prediction [73]. Nevertheless, MLPs suffer from some common drawbacks including network's complexity, numerous network variables required to be optimised and random weight initialisation.

Another ANN architecture that has been found suitable for numerous applications in recent years is the PNN. The PNN, first introduced by Specht in 1988 and 1990, is capable of defining decision boundaries exploiting Bayesian theory [23, 24]. This quality makes the PNN an effective technique for classification and pattern recognition tasks, especially when used for datasets containing samples adequately correlated to the associated class and well separated categories [25]. The PNN also trains faster than the MLP and requires fewer variables to be optimised. Although this method has been known for quite a while, it has only recently been identified as an effective classifier for cancer diagnosis and prediction [26, 27].

2.6.1. Advantages

The basic requirement of supervised ANNs for data classification is a large enough dataset with known inputs and outputs for training and test. In this, ANNs are similar to any other statistical or machine learning algorithms employed as a DSS. ANNs owe their extensive application in cancer prediction to several special characteristics in their theory and structure. These include ANNs' nonlinear modelling and maintaining adaptable connection weights which makes them suitable for data driven analysis without making any prior assumptions about the underlying models.

ANNs can achieve good results with fewer data compared to data sizes estimated by statistical sample size computational methods [43]. It is because the ANNs are efficient in using the available data by extracting higher-order statistics. This also gives flexibility to ANNs in achieving significant results in a wide variety of applications and especially in medical diagnosis [68].

2.7. Neural network performance evaluation

One of the important issues in ANN classification is whether the approximated classification rule based on limited sets of data can represent a whole population of patients. Therefore, the designed ANNs need to be assessed with new data which is not used for constructing the classification rule in the training procedure.

In recent years, many prognostic models have been developed to perform the task of medical diagnosis and classification using ANNs. An extensive consideration has been given to the stage of classifier design for developing the prognostic model. Conversely, the stage of the classifier assessment given the limited available dataset has been overlooked by many studies. This issue was raised in 1995 and the importance of validating prognostic models in order to ensure the effectiveness of the models was highlighted [74].

To assess the ability of the network in predicting the outcome, the network predictive performance for new patients should be evaluated. For this purpose, the available data are divided into two parts: the training and the test sets. The designed network is trained using a training set of data. The training set *X* contains the markers and the known output, both of which are used to facilitate the learning process via supervised learning. The test set is kept unseen and used only after training is complete to estimate the performance of the ANN. The implemented ANN must provide a set of classifications for the records in the test set which are then being compared with the known output to deduce the accuracy of the system. For a dichotomous target $t_{i,}$ the network provides dichotomous outputs o_i and hence, the output error E_i would take only two values as

$$E_i = \mathrm{E}(o_i, t_i) = \begin{cases} 0 & if \quad o_i = t_i \\ 1 & if \quad o_i \neq t_i \end{cases} \tag{2.35}$$

An ideal error estimation method is the one with zero bias and variance. However, achieving such estimation is unfeasible in practice as the dataset is only a small sample of the population and hence a trade off between the bias and variance of the error estimation technique is sought.

2.7.1. Estimation of classification error

ANNs classify new patterns by constructing prediction rules from a training dataset. After training the ANN with teacher, it is necessary to measure the accuracy of the designed network. The simplest measure for the network accuracy is the training error also denoted as apparent error $\boldsymbol{E^a}$in some texts. This is performed by testing the classifier with the same data employed for training and is defined as the ratio of the misclassified samples to the total number of samples when the training is stopped:

$$\boldsymbol{E^a = \frac{1}{N}\sum_{i=1}^{N} E(o_i, t_i) = \frac{1}{N}\sum_{i=1}^{N} E_i} \tag{2.36}$$

This measure is not a true presentation of the classification accuracy. This is because the prediction rules are constructed using the same training patterns and hence provide an error estimation that is significantly biased downward (i.e. overoptimistic). Therefore, the error rate obtained by testing the network with the same training set is not a reliable measure of the network's error rate for new patterns. This is a well known fact and medical studies using ANN for diagnosis do not exploit training error as a measure of classification error [20].

In order to reduce this bias, holdout method is carried out by random division of data into training and test sets where the classifier is designed using merely the training data and hence the test data are kept unseen. Holdout method is commonly used in experiments carried out with ANNs. However, it is not an efficient method in small datasets since a part of data are held for test and it cannot be used for the network training. Holdout method would also give a high variance in the estimated error rate depending on the random selection of training and test data.

Another method preferred over holdout is *k*-fold CV. In this method, data is divided into *k* groups. Then, the network is trained with *k-1* groups keeping one group as the test set. This procedure is repeated for *k* times until each group is used once for test. Final prediction accuracy is then computed as the average of *k* test results. Hence, all data are used for training the network at least once and the test set is kept unseen to the network. Furthermore, CV is known to provide an unbiased estimate of error rate. However, this low bias comes at the cost of high variability especially for discontinuous outputs for which *k*-fold CV error estimate function is a discontinuous function of training set [75].

Bootstrap technique is an alternative method for CV which reduces the variability in error estimate by using data resampling technique and maintains a low bias in estimating the misclassification error [76]. This method is detailed in chapter 4 as the main technique employed to evaluate the designed ANNs in this study.

2.8. Summary

In this chapter, a review of the foundation of medical DSSs has been explained. Then, a brief history and principles of the most widely employed statistical and machine learning methods for breast cancer prediction including LR, DA and DT was included. The discussion on the advantages and disadvantages of these methods has demonstrated the ineffectiveness of these methods in medical diagnosis. Regression models are the most commonly used statistical method in medical studies which make no assumptions about the underlying distribution of the independent variables. However, they have several drawbacks including their limitation to linear modelling and their requirement for independent variables and a large sample size. The disadvantages of LDA that hinders its widespread application in medical studies include its limitation to continuous input variables and input data assumptions such as multivariate normality and equal covariance for independent variables. Among various machine learning techniques, DTs are the most widely used methods in cancer diagnosis. However, they are prone to overfitting and are limited to variables with small number of discrete values. In general, the relationship between multiple predictors is typically a nonlinear and complex one which renders the linear statistical methods insufficient for these classification problems.

The sections on the statistical and machine learning methods have led to the conclusion that a more effective classifier is required for capturing the nonlinearity in the medical data. ANNs were then introduced as valuable classification tools that maintain several advantages over conventional statistical methods especially for data that exhibits nonlinearity. This chapter has also covered the principles of ANNs and their history in medical diagnosis. In addition, the necessity for reliable error estimation and the currently employed methods were detailed in this chapter. The main ANN architecture and evaluation method employed in this study are explained in chapter 4.

Chapter 3

3. Minimally invasive biomarkers for breast cancer prediction

3.1. Introduction

In recent year, many studies have focused on devising an approach to avoid the invasive procedures for predicting breast cancer. Various tumour biomarkers have been discovered and tested in this regard to enable physicians to predict breast cancer before performing invasive treatments. However, individual tumour markers alone are not a good predictor of breast cancer progression as a combination of various biomarkers can give a better prediction of nodal status in individual patients. In this chapter, a novel set of biomarkers is proposed for the purpose of predicting the progression of breast carcinoma. While these markers have been known in the literature of breast cancer prediction, their combination together forms a novel relation with the output of breast cancer progression considered in this study.

The first section of this chapter, introduces the measure employed for quantifying breast cancer progression. The biomarkers chosen in this study are then listed and the dataset characteristics, in addition, to the biomarkers' measurement methods and descriptive statistics are explained. The third part of the chapter entails a medical literature of the chosen biomarkers to establish their individual roles and application in predicting breast cancer progression. In the final section, the scatter plots of the markers with regards to the breast cancer output are illustrated and their correlation coefficients are presented. This section aims to investigate the relationship between different markers and the output visually and quantitatively.

3.2. Breast cancer prediction using axillary lymph node metastasis

Deciding the state of tumour spread to the ALNs is the most influential prognostic factor in women diagnosed with breast cancer [77]. The status of ALN involvement is an important factor in making the appropriate decision about treatment type and timing for patients with breast carcinoma. The invasion of cancer to the ALNs is typically determined by excision of ALNs and histological detection of tumour cells in the specimens, or by SLNB. ALN dissection and SLNB are invasive procedures and may cause pain, bleeding and infection. Besides, they are costly and patients diagnosed with no invasion to the ALNs after node dissection and the SLNB are imposed with unnecessary morbidity. Therefore, it is important to devise minimally invasive surrogate methods to determine the potential presence of any metastasis to the ALNs prior to surgery.

In recent years, much attention has been focused on the prediction of ALNs using various tumour markers such as clinical, pathological and radiological features. Patani et al. [7] identified these markers as predictive and non-prognostic, since predictive factors are employed for predicting metastasis to the ALNs, whilst prognostic markers are employed for predicting clinical outcome such as overall survival. There is little agreement to date on single or a combination of biomarkers for accurate nodal status prediction that can replace the conventional techniques such as axillary dissection or SLNB. Single tumour biomarkers alone do not predict the potential progression of breast cancer to ALNs. A combination of various biomarkers can give a better prediction of nodal status in individual patients.

3.3. Breast cancer dataset

In an attempt to find minimally invasive surrogate techniques for predicting the spread of breast cancer to the ALNs, this study has identified several molecular and cellular markers that dominate the field of tumour progression. These can be classed as: (a) cell cycle regulatory genes and cell proliferation markers, (b) metastasis promoter and suppressor genes, (c) hormonal and growth factor receptors, and (d) biological features. The expression of molecular markers is measured in tumour tissue using immunohistochemistry (IHC)[2] to quantify the protein. These parameters

[2] IHC is the procedure of detecting antigens in cells based on the binding of antibodies to particular antigens in biological tissues.

are routinely measured on fine needle aspirate (FNA) cells derived from the primary tumour.

The dataset employed in this study was collected in the Department of Clinical Oncology and Department of Anatomy, Pathology and Histology, Infermi Hospital, Rimini, Italy. The data consist of records of 108 patients from which 47 patients were diagnosed with metastasis in ALNs (node positive) and 61 did not show any metastasis (node negative). The data include information about the five tumour markers viz. tumour size, ER, PR, p53 and Ki-67, in addition to the age of each patient at the time of diagnosis (aged 27 to 83 years old).

3.3.1. Measurements and descriptive statistics

Tumour size was determined by measuring the largest dimensions of the tumour from mammograms[3] and was expressed in centimetres as a continuous variable with a range of 0.5 to 5.5 centimetres. ER and PR levels were simply expressed as *present* (1) or *not present* (-1). (1) was denoted when the receptors were detected by IHC in more than 10% of cells or when receptor level was greater than 20 fmol/mg protein while (-1), when receptor levels were below 20 fmol/mg protein[4]. The expression of p53 and Ki-67 was given as a percentage of cells stained positive in IHC tests and was stated as continuous variables. The presence of nodal metastases was also available for each patient and has been considered as the system output. Nodal data

[3] Because the tumour does not have a regular shape, the greatest dimension is used as a indication of size.

[4] fmol/mg protein is a a way referring to the amount of ER or PR receptor protein in one mg of total cell protein; fmol is an abbreviation for femtomole which is 10 to the power of -15 of a mole (i.e. a billionth of a mole). Hence, receptor level of fmol/mg protein refers to femtomole of receptor /gram of cell protein.

were originally provided as the number of positive nodes over the total number of nodes tested. To employ nodal status as ANN output, node negative patients were designated as (-1) and patients with even one positive node were designated as (1). Markers considered in this study and their descriptive statistics are tabulated in Table 3.1 and Table 3.2 for continuous and binary markers respectively. A description and history of these biomarkers in clinical domain is provided in the following sections.

Table 3.1: Continuous tumour markers and their descriptive statistics

Tumour marker	Minimum	Maximum	Range	Mean	Std. deviation
Tumour Size (cm)	0.50	5.50	5.00	2.07	0.92
p53 (%)	0.00	0.90	0.90	0.16	0.26
Ki-67 (%)	0.03	0.80	0.77	0.31	0.16
Age (years)	27	83	56	58.12	12.88

Table 3.2: Binary tumour markers and their descriptive statistics

Tumour marker	Frequency of 1	Frequency of -1
ER (Oestrogen Receptor Status)	69%	31%
PR (Progesterone Receptor Status)	56%	44%

3.4. Biomarker descriptions

3.4.1. Tumour size

Tumour size has been identified as a significant prognostic marker both independently and in combination with other markers for ALN prediction [78]. Progressive tumour growth increases the likelihood of vascular and lymphatic dissemination and so the size of the tumour is directly related to higher chances of local metastasis, recurrence and death. It also has an influence on the prognosis of both node-positive and node-negative cases, where increased tumour size has been associated with decreased survival regardless of lymph node status [79]. However, this effect is greater in node-positive than node-negative patients as reported by Carter et al. [80]; whereas, no relation between tumour size and survival was observed in node-negative patients by Vallagussa et al. [81]. Nevertheless, size of the tumour is considered as one of the most significant predictors of tumour behaviour. It has been reported that patients with small tumours (less than or equal to 1.0 cm) have significantly higher rates of relapse-free survival than those with larger tumours, with 96% remaining relapse free at 5 years.

3.4.2. Oestrogen and progesterone receptor status

The female steroid hormones, oestrogen and progesterone, stimulate the growth of a variety of target tissues via binding to their respective receptors (ER and PR), which promotes tumor growth. The hormone-receptor complex functions as a transcription factor and initiate the transcription of responsive genes resulting in appropriate physiological function. Oestrogen and progesterone are known to influence the function of different homeobox transcription factors in breast cancer cells. Therefore, they could be inducing the transcription of specific target genes associated with cell proliferation and differentiation.

The presence of steroid hormone receptors subserves an important function as effective therapeutic targets. Breast cancers that express oestrogen and progesterone receptors are targeted by using anti-oestrogens. Hence, the absence of ER in breast cancer is regarded as an indicator of poor prognosis, since ER- tumours would be resistant to anti-oestrogen therapy, and can continue to grow rapidly and result in poor patient outcome. Hypermethylation and down-regulation of ER gene expression also occurs in many human tumours. In some cases, the silencing of the ER gene correlates with disease progression. ER and PR expression status is significant in determining response to endocrine therapy. Patients expressing both receptors have the best prognosis and are more likely to respond to hormone treatment than patients with ER+/PR- tumours [82]. PR+ patients however, appear to respond better to hormonal therapies with a higher survival rate consequently [83].

In breast carcinoma, the PR gene is also down regulated. PR seems to be the main component of ER functional pathway and ER could be regulating PR. In some instances, ER+ breast cancers have been found to be refractory to anti-oestrogen therapy. ER are known to be able to induce the expression of PR, and often a reduced response to hormones is perceived as a non-functional state of ER [84]. Hence, PR has been recognized as a prognostic factor independent of ER. In employing these markers for assessing tumour progression, one should note the interaction between the signal transduction pathways of these steroids. In breast cancer cells, both oestrogens and progesterone can activate a common signalling pathway viz. the Src/Erk pathway. Ballaré et al. [85] have attributed this to the two domains of PR which interact with ER. On the other hand, progesterone can negatively regulate other oestrogen regulated signalling pathways leading to inhibition of proliferation [86]. In other words, cross talk between steroid receptors is bound to impinge upon progression of the disease.

3.4.3. p53

Only around 20% of all human breast cancers have a *p53* mutation [87], but specific types of breast carcinomas such as medullary breast cancers have higher p53 mutation frequency (100%) [88]. Mutations in the *p53* gene or inactivation of its protein has been shown to play a role in several types of cancers [89]. *p53* functions as a cell cycle regulator. In cells with damaged DNA, there is an over expression of *p53* and cells are prevented from entering the synthetic phase until the DNA damage is repaired. *p53* may also trigger a cell suicide response in cases where DNA is irreparable. Where *p53* is mutated or lost, the cells continue to divide and replicate the damaged DNA, passing on mutations to daughter cells. *p53* is known to prevent the formation of tumours by preventing progression of the cell cycle, in particular G1-S and G2-M transitions. Therefore, *p53* is considered as a good guide for the prediction of breast cancer patient survival. Nevertheless, there is little evidence that *p53* can be used as an indicator of initial tumour formation [90]. The perceived role of *p53* in cell cycle regulation is not often considered significant, as most studies recognize that there is a poor correlation between *p53* and metastasis in breast cancer. As an instance, Erdem et al. [91] suggested that *p53* offers no significant prognostic value.

Studies have shown that breast cancer patients with a mutant *p53* respond differently to those with wild-type by being resistant to certain chemotherapy regimens while sensitive to others [92]. However, it is still not possible to conclude how *p53* mutations can impact breast cancer therapies and patient's overall clinical outcome which in turn highlights how several other factors might be influencing the *p53* network [93].

3.4.4. Ki-67

The Ki-67 protein is a growth promoter antigen, present in all but the resting G0 cells. The Ki-67 antigen is expressed on the outer surface of the nucleolus, especially in its granular component during late G1, G2 and M phases of the cell cycle. So the determination of its expression serves as an accurate cell proliferation index and hence employed in a diagnostic and prognostic role [94]. It has been shown that evaluation of Ki-67 antigen can serve as the proliferative determinant of early breast cancer as reviewed by Azambuja et al. [95]. Even though, Ki-67 labelling index which is a percentage of Ki-67 antigens detected using nuclear immunoassays is determined by histopathologists, it cannot be used as an accurate clinical prognostic marker for breast cancer. Ki-67 labelling index can vary from one histopathologist to another depending on the site of tumour selection for counting, cut-off levels, as well as the detection kits or antibody used for the immunoassays [96].

In a recent study, it has been demonstrated that a combination of biomarkers which comprises of both *Ki-67* and *p53* are able to better predict the outcome of early ER+ breast cancer patients [97]. Thus, the wide spectrum of molecular markers has afforded an excellent opportunity to relate their expression as an integrated body of biological response modifiers to breast cancer progression. However, their apparent complex interaction has necessitated the development, evolution and validation of artificial intelligence techniques such as ANN. These techniques offer the possibility of assessing various markers and their biological effects for their ability to predict the cancer progression so that appropriate treatments can be tailored to combat the disease.

3.4.5. Age

As a very basic part of the patient dataset, age is still included in most analyses of breast cancer. However, the relationship is not necessarily that simple because aggressive tumours may form early in life cycle. But, age is commonly used in deciding treatment levels as it can be considered as an indicator of the resilience of the patient to aggressive treatments.

Early breast cancer incidents are increasing in young women. It has been frequently reported that women who are younger than 35-years old have a poorer prognosis than women older than 65-years old. Patients less than 35 years of age have significantly shorter survival time which is also true in the presence of other good prognostic factors and regardless of their menopausal status. The underlying reason however, remains unclear. Younger women diagnosed with breast cancer are more likely to have ALN metastasis, higher frequency of undifferentiated tumours, with high histopathological grading while also negative for ER. In addition, younger patients have a higher risk of local recurrence and developing distant metastases. In recurrent breast cancer, one study predicted the median survival times after the first recurrence to be 491 days for patients less than 35 years of age, 590 days for patients

36 to 45 years of age, and 700 days for those greater than 45 years of age [98]. Moreover, metastasis-free survival and overall survival are also significantly shorter in younger age groups. All this indicates that breast cancer is biologically more aggressive among younger women compared to older women and therefore, young age has an unfavourable impact on prognosis in breast cancer.

3.5. Scatter plots and correlation coefficient

In order to inspect the relationship between different tumour biomarkers and the output (state of nodal involvement), data scatter plots for each two pairs of inputs are plotted. These scatter plots are good indicators of potential associations between each two pairs of markers and the degree of output separability based on each set of two biomarkers.

Data scatter plots for all the two group combinations of input variables with regard to data scatter plots for all the two group combinations of input variables with regard to their corresponding output are illustrated in Figures 3.1 and 3.2. In Figure 3.1, the marker combinations including only continuous input variables are illustrated, while in Figure 3.2, at least one of the biomarkers have discrete values. The axis labels in each subfigure show the corresponding marker. In addition, output refers to the state of the nodal involvement where outputs with a value of *1* are indicated by circles ‘x’ while outputs equal to *0* are indicated by cross signs ‘o’. The continuous variables illustrated in these figures are standardised with regard to their mean and standard deviation by subtracting the vector mean from the vector and dividing it by its standard deviation. Therefore, each continuous variable maintains a zero mean and a unit standard deviation.

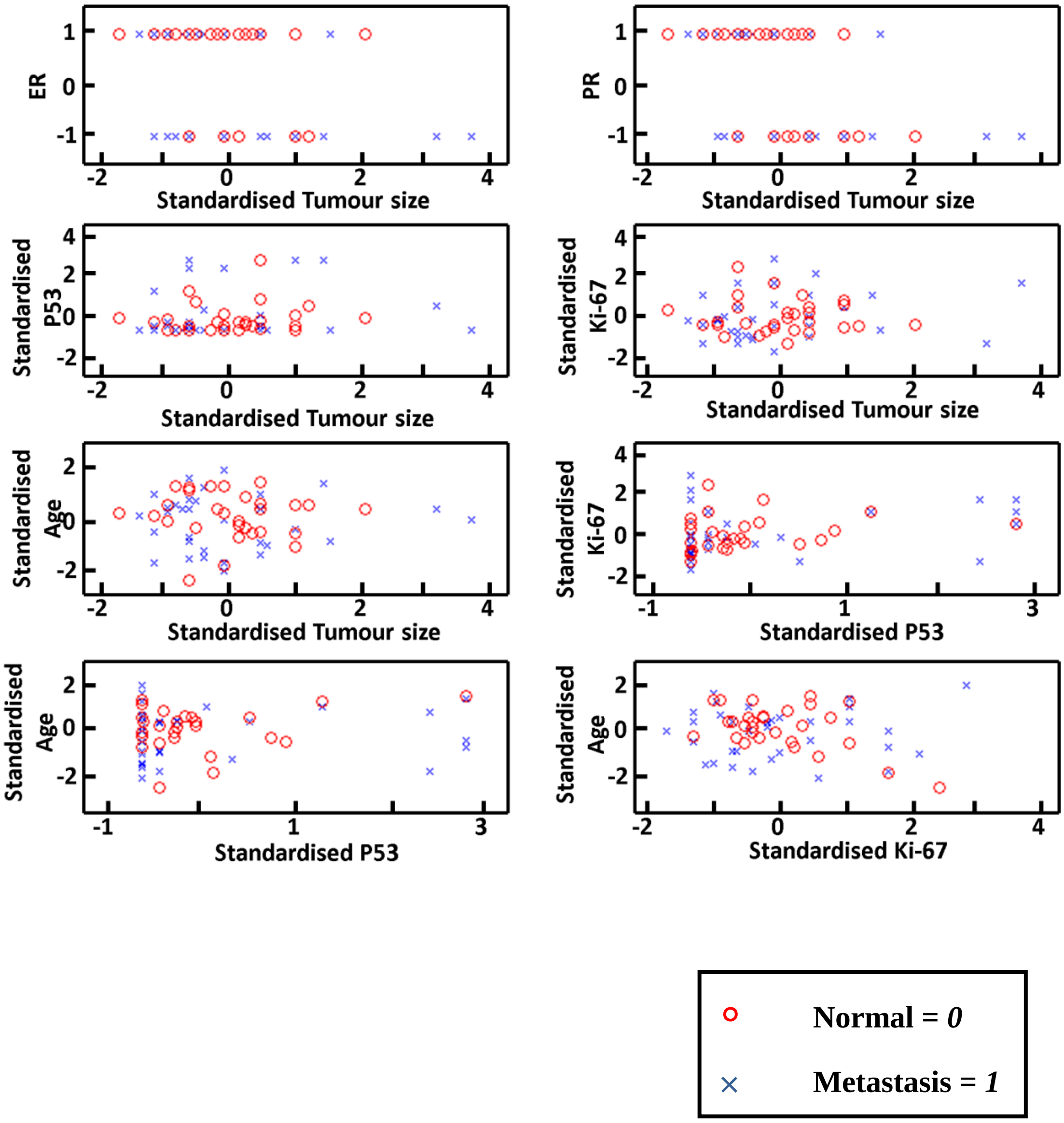

Figure 3.1: Data scatter plots for all two group combinations of continuous input variables with regard to their corresponding output where the output refers to the state of nodal involvement. The continuous variables are standardised such that they maintain zero mean and unit standard deviation.

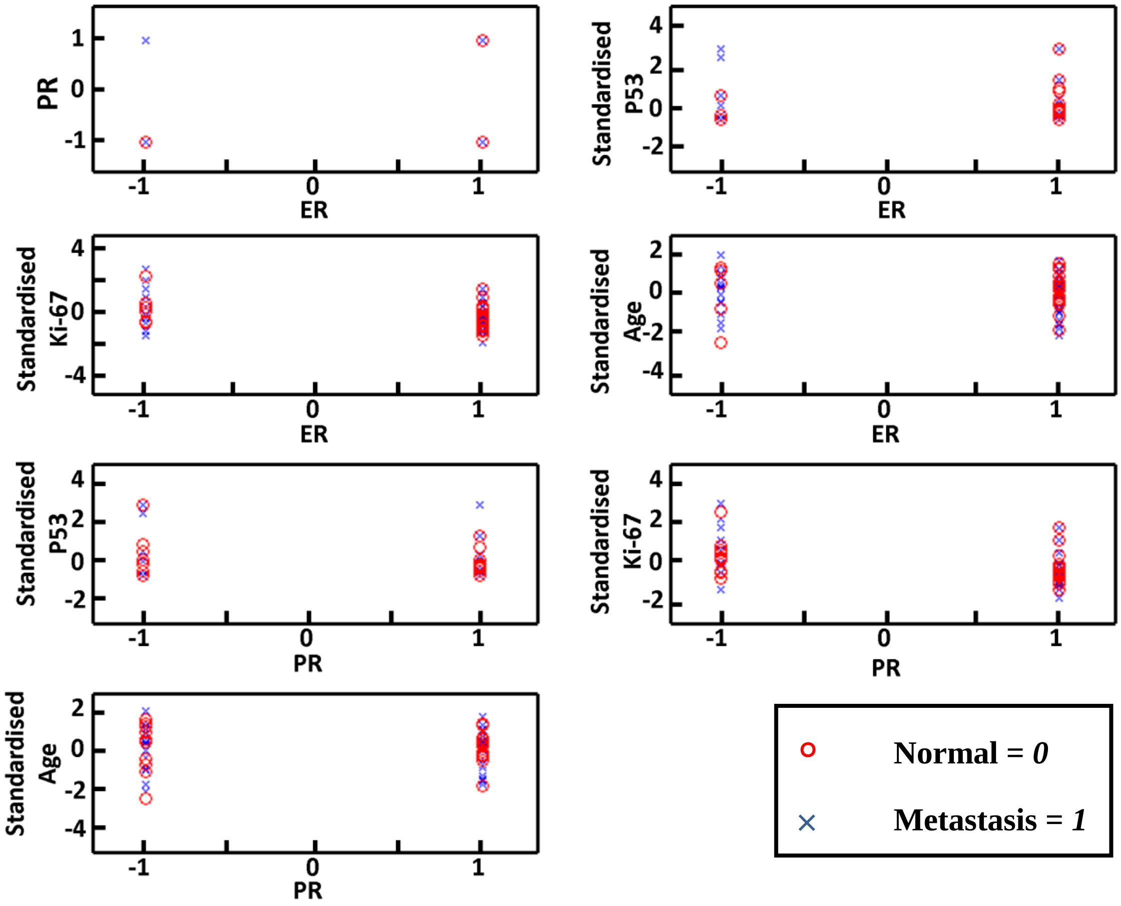

Figure 3.2: Data scatter plots for two group combinations of input variables with regard to their corresponding output where at least one of the biomarkers have discrete values. The output refers to the state of nodal involvement. The continuous variables are standardised such that they maintain zero mean and unit standard deviation.

The scatter plots of the two-marker combinations in figures 3.1 and 3.2 illustrate the separability of output based on each set of biomarkers. From these figures, it is clear that the state of nodal involvement is not linearly separable with any of the 2-marker combinations. The complex nature of the marker relationship in the 2-dimensional figure is also evident from this figure.

In order to evaluate these marker relationships numerically, the PCC explained in section 2.2.2 is computed for all the tumour biomarkers and the state of nodal involvement. These results are tabulated in Table 3.3. The PCC presented in this table takes a number between -1 to +1 for the measure of linear dependence between tumour biomarkers. A positive value represents a positive linear relationship while a negative one implies negative linear relationship and 0 suggests no linear relation between variables. The diagonal cells of the table have a value of 1 since they show the correlation of the tumour biomarker with itself. Additionally, this table represents a symmetrical matrix for PCC since for two variables A and B, $PCC(A, B) = PCC(B, A)$.

Table 3.3: The Pearson correlation coefficient computed for all the tumour biomarkers and the state of nodal involvement

	Tumour size	ER	PR	p53	Ki-67	Age	Nodal involvement
Tumour size	1	-0.288	-0.472	0.109	0.111	-0.008	0.026
ER	-0.288	1	0.615	-0.265	-0.161	-0.058	0.202
PR	-0.472	0.615	1	-0.370	-0.214	-0.080	0.041
P53	0.109	-0.265	-0.370	1	0.278	-0.141	0.011
Ki-67	0.111	-0.161	-0.214	0.278	1	0.079	-0.079
Age	-0.008	-0.058	-0.080	-0.141	0.079	1	0.127
Nodal involvement	0.026	0.202	0.041	0.011	-0.079	0.127	1

Results from Table 3.3 suggest some linear relation between ER and PR (PCC = 0.615). The degree of linear dependence of P53 and PR (PCC = -0.370) and tumour size and PR (PCC = -0.472) is also noticeable. From the correlation between the nodal involvement and other tumour markers indicated in the last row of the table, it is evident that there is no significant linear relation between markers and the output. These results however, do not necessarily provide any indication about the existence of any nonlinear interaction between the different markers and the output.

From Figure 3.1 and Table 3.3, it is clear that there is little or no linear relation between the tumour biomarkers and the state of nodal involvement. This signifies the inefficiency of linear statistical classifiers such as LR in classifying the breast cancer data and suggests the need for nonlinear classifiers such as ANNs.

3.6. Summary

In this chapter, the ALN status for predicting the breast cancer progression was explained and a novel set of markers for predicting the outcome was introduced. Although the presented markers have existed in breast cancer literature for a long time, this research proposes a novel relation between these markers and the state of ALN progression.

The relationship between the presented biomarkers and breast cancer progression outcome was investigated from three different aspects including medical, visual and quantitative analysis. The medical aspect was covered in a medical literature about each individual marker. This also entailed a description of the biomarkers and their role in breast cancer progression. To show this relationship visually, scatter plots of the data were presented. These scatter plots illustrate the complex nature of marker relationships and the inseparability of the markers with respect to the cancer outcome in 2-dimensional space. Although it is not possible to visualise this relationship in higher dimensional spaces, it is expected that in higher dimensional spaces, the markers provide a higher degree of separability for patients with and without metastasis. To quantify the relation of markers and the outcome, the PCC is employed. PCC represents the degree of linear relationship between each two markers or each marker and the outcome. These results show little or no linear relation between the markers and the state of ALN progression.

In the next chapter, an ANN architecture and evaluation method are proposed to predict the state of metastasis to the ALNs using the chosen biomarkers. These methods are chosen in combination to overcome some common problems of the commonly employed methods detailed in chapter 2. In addition, these methods are chosen such that they address the requirements for classifying the ALN progression status using the chosen set of biomarkers.

Chapter 4

4. A neural network approach for reliable breast cancer prediction

4.1. Introduction

As mentioned in chapter 2, some common problems exist in the most widely employed statistical and machine learning studies conducted in the field of cancer diagnosis. These include limitation of linear modelling, large data size requirement and several assumptions made about the input data. To address these common problems, this research exploits an ANN structure known as PNN. The PNN is employed to predict the state of metastasis to the ALNs using a breast cancer dataset from 108 patients. This dataset and the patient biomarkers employed to predict the nodal status have been explained in detail in the previous chapter.

This chapter details the MLP structure as the most widely employed ANN in cancer studies, here used as the benchmark for the comparison of results. MLPs disadvantages and common issues in cancer studies are then covered. After that, an in depth explanation of the PNN; its structure and training algorithm are detailed. The final section of the chapter describes an error estimation method based on resampling, entitled as .632 bootstrap. The applied methods in conjunction with the employed dataset bring about a novel application of PNN in breast cancer prediction.

4.2. Multilayer perceptron

The MLP is a feed-forward ANN in which the data, entered at the input layer, propagate in one direction through the hidden layer(s) to the output layer (Figure 4.1). Hidden and output layers are composed of single units called perceptrons. Each perceptron in the MLP receives a set of inputs which are first weighted and then added together. The resultant value is used to trigger an activation function that will map the combined inputs to an appropriate output response.

In essence, the MLP can be deemed as a simple input-output model with connection weights as the free parameters. Such a model is able to represent a function describing the relationship between the inputs and outputs where the number of hidden layers and number of units in each layer can define the complexity of the function. Therefore, one critical issue in MLPs is defining an optimum number of hidden layers and their corresponding units such that the constructed function is complex enough to describe the relation between the inputs and outputs but not too complex, so that it maintains the ability to generalise for new data [62, 65]. The number of neurons in the input and output layers is defined by the problem in hand. In the input layer, this is equal to the input space dimension while in the output layer, it is defined by the output categories.

Learning of the input-output relationship in MLPs is performed by the adjustment of the connection weights linking the layers together. This is done via a training algorithm that is a predictive model which adjusts the weights by minimising the network error with respect to its weights. Back propagation (BP) [62] is the most widely implemented training algorithm applied for MLP training that is detailed in the next section.

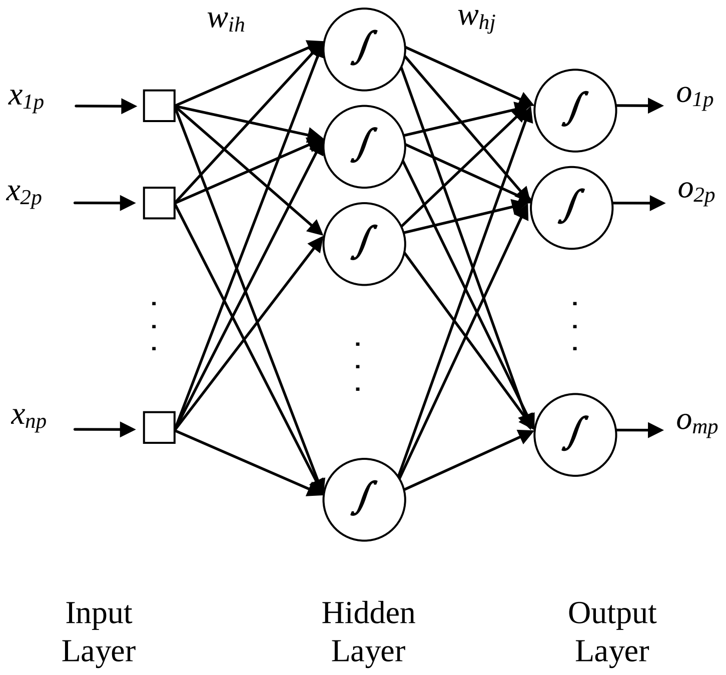

Figure 4.1: Structure of the MLP with one hidden layer

4.2.1. Back Propagation (BP)

BP algorithm is employed to train the MLP using a sample set of data so that the network can predict a new input correctly. The principles of BP training are similar to perceptron's training explained in 2.5.3. Generally speaking, this is performed by initialising the network weights randomly and computing the network output using the available input data. This output is then compared with the desired outcome and an error signal is computed which is then employed to construct an error function. Finally, the error function is minimised with respect to weights to adjust them such that the network provides a closer output to the desired outcome. This procedure is replicated several times, each round called an iteration, to obtain the optimum set of weights.

BP algorithm consists of two passes of information: a forward and backward step. In the forward step, the predicted outputs O_{jp} corresponding to the given inputs of the layer y_h for the p^{th} input pattern are evaluated using:

$$o_{jp}(\mathrm{s}) = \varphi(\sum_{h=0}^{l} w_{hj}\,(\mathrm{s})y_{hp}(\mathrm{s})) \tag{4.1}$$

where s denotes the s^{th} iteration (step) of the training process y_h is a vector of outputs of the hidden layer for the p^{th} input pattern, computed as:

$$y_{hp}(\mathrm{s}) = \varphi(\sum_{i=0}^{n} w_{ih}\,(\mathrm{s})x_{ip}(\mathrm{s})) \tag{4.2}$$

where x_{ip} is the i^{th} element of the p^{th} input pattern, and w_{ih} and w_{hj} are the connection weights linking the input-hidden and the hidden-output layers respectively. It should be noted that w_{h0} and w_{0j} are equal to 1 and correspond to each neuron's bias with their associated inputs being equal to the bias value. The function φ_h is the nonlinear activation function (transfer function), commonly considered as a hyperbolic tangent sigmoid function in both hidden and output layers and hence represented with the same symbol for both layers:

$$\varphi(y) = \frac{e^{2y} - 1}{e^{2y} + 1} \qquad (4.3)$$

As mentioned in 2.5.2, this function is used for its special characteristics such as monotonicity and symmetrical output. When the forward pass is completed, an error signal e is computed by comparing the network's outputs o_j with the desired response t_j. The error signal is then fed into a cost function to control the adjustments to the weights. Mean square error (MSE) is commonly used as the cost function which is drawn from maximising the likelihood of each data point (x_p, t_p) in the training data defined as:

$$\xi = \frac{1}{2Nm} \sum_{p=1}^{N} \sum_{j=1}^{m} \left(t_{jp}(\mathrm{s}) - o_{jp}(\mathrm{s})\right)^2 \qquad (4.4)$$

where the cost function ξ is the mean of squared-error of the total number of output neurons m over the total number of training patterns denoted by N . t_{jp} and o_{jp} are the desired output and the network's output respectively, resulting from the j^{th} output neuron using the p^{th} input pattern.

In the backward pass, gradient descent (GD) algorithm is used for updating weights. In this method, the partial derivative of the cost function with respect to network weights is computed and the weights are adjusted accordingly. To explain this procedure clearly, one neuron from Figure 4.1 is considered and the amount of weight adjustment is computed for it. This can be easily generalised for the whole network. For simplicity, this neuron and its connections are illustrated in Figure 4.2.

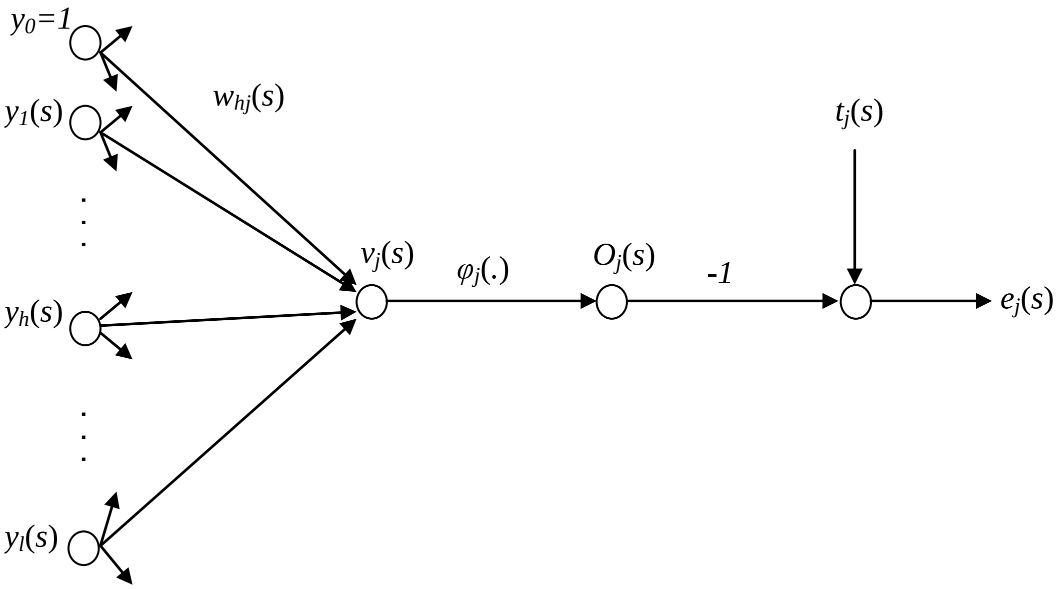

Figure 4.2: An output neuron *j* and its connections from a multilayer perceptron

As illustrated in Figure 4.2, neuron *j* receives the output signals of the previous layer's neurons. These signals are weighted to produce the activation of the neuron *j* as:

$$v_j(s) = \sum_{h=0}^{l} w_{hj}(s) y_h(s) \qquad (4.5)$$

where *l* is the number of inputs to neuron *j*. The first input $y_0=1$ is the bias applied to neuron *j*. the neuron's output $o_j(s)$ is computed by feeding the activation signal $v_j(s)$ into the transfer function φ:

$$o_j(s) = \varphi_j(v_j(s)) \qquad (4.6)$$

In each iteration, the weights are modified by $\Delta w_{jh}(s)$ that is proportional to the partial derivative $\frac{\partial \xi}{\partial w_{jh}(s)}$. Using he chain rule, this can be expressed as:

$$\frac{\partial \xi}{\partial w_{jh}(s)} = \frac{\partial \xi}{\partial e_j(s)} . \frac{\partial e_j(s)}{\partial o_j(s)} . \frac{\partial o_j(s)}{\partial v_j(s)} . \frac{\partial v_j(s)}{\partial w_{hj}(s)} \qquad (4.7)$$

Using the general form of the MSE cost function in (4.4) as $\xi = \frac{1}{2} \sum e_j^2(s)$, this leads to:

$$\frac{\partial \xi}{\partial e_j(s)} = e_j(s) \qquad (4.8)$$

The second term in the right hand side of (4.7) can be obtained from $e_j = t_j - o_j$ as $\frac{\partial e_j(s)}{\partial o_j(s)} = -1$.

The third term in the right hand side of (4.7) can be obtained using (2.33) as:

$$\frac{\partial o_j(s)}{\partial v_j(s)} = \varphi_j^{/}\left(v_j(s)\right) \tag{4.9}$$

Finally, differentiating (4.5) with respect to the weights gives:

$$\frac{\partial v_j(s)}{\partial w_{hj}(s)} = y_h(s) \tag{4.10}$$

Using equations (4.8) to (4.10), the equation (4.7) can be written as:

$$\frac{\partial \xi}{\partial w_{hj}(s)} = -e_j(s)\varphi_j^{/}\left(v_j(s)\right)y_h(s) \tag{4.11}$$

The weight change is defined proportional to the partial derivative in (4.11) as:

$$\Delta w_{jh}(s) = -\beta\frac{\partial \xi}{\partial w_{hj}(s)} = \beta\delta_j(n)y_h(s) \tag{4.12}$$

where β is the learning rate (step size) in BP and the local gradient $\delta_j(n)$ is defined as:

$$\delta_j(s) = \frac{\partial \xi}{\partial v_j(s)} = \frac{\partial \xi}{\partial e_j(s)}.\frac{\partial e_j(s)}{\partial o_j(s)}.\frac{\partial o_j(s)}{\partial v_j(s)} = -e_j(s)\varphi_j^{/}\left(v_j(s)\right) \tag{4.13}$$

The local gradient is a vector pointing along the steepest descent from the present point. A series of such steps (changing the weights in the direction of the local gradient) guarantees declining of the error and moving towards the minimum of the error surface. β defines the size of the steps in the direction of the steepest descent. While, a large β helps the algorithm to converge quickly towards the minimum, it might cause overstepping the answer. On the contrary, a small step may lead the algorithm to the solution, but it may require a long time and many iterations.

Therefore, an appropriate choice of the step size is important in achieving a desirable solution within a reasonable amount of time. This depends on the application in hand and is usually selected by experiment. It is also possible to decrease the step size as the algorithm progresses towards the minimum [62].

In general, the training algorithm progresses towards the solution in each iteration by feeding the inputs to the network and computing an error by comparing the network output and the targets. The weights are then adjusted according to the error function. The training stops when a minimum error is reached or a certain number of iterations are passed. These are called the stopping criteria which prevent the algorithm to proceed for a long time and overtrain.

Although BP is widely employed for training the MLP, it has some limitations. The main limitations are the slow convergence of the BP and the high probability of the algorithm getting trapped in local minima on the error surface (the surface obtained from the cost function). These limitations can be obviated by using the second derivative of the error function in the scaled conjugate gradient (SCG) algorithm.

4.2.2. Scaled conjugate gradient algorithm

SCG algorithm, proposed by Moller [99], is an effective approach to perform supervised learning in feed-forward MLPs. SCG algorithm is a class of conjugate gradient optimisation techniques which is much faster than usual BP algorithms. The algorithm consists of a forward and backward step. The forward step is similar to BP and has been covered in the previous section.

In the backward pass, second partial derivatives of the cost function with respect to the network parameters, weights and biases, are computed and propagated back through the network. Second derivative of the MSE function in (4.4) with respect to the weight vector W is denoted as Hessian matrix H, computed as:

$$H(s) = \frac{\delta^2 \xi(W)}{\delta W^2} \tag{4.14}$$

where W denotes the network's weight matrix obtained from concatenating the two layers of weights w_{ih} and w_{hj}. Using the second order derivative of the cost function provides additional information related to the curvature of the MSE cost function (error surface), and hence, results in faster and more accurate convergence to the minimum point compared to first order techniques, such as the standard BP, that use the first derivatives only.

However, in conjugate gradient algorithm this is done at the cost of a high computational complexity as it performs a line-search to minimise the cost function. A line-search involves defining a search direction in the weight space and locating the minimum of the cost function along that direction [65]. In conjugate gradient algorithm, the weights at each layer are updated using the following recursion:

$$w_{s+1} = w_s + \beta_s d_s \tag{4.15}$$

where d_s is the search direction at step s which is defined such that the component of the gradient parallel to the previous search direction is kept equal to zero, and hence, search directions are described as being conjugate. β_s defines the step size in each step which can be derived as:

$$\beta_s = -\frac{d_s^T g_s}{d_s^T H d_s} \tag{4.16}$$

where g_s is the gradient vector of the error surface at step s. From (4.16), it can be inferred that Hessian matrix H needs to be evaluated in each step in order to define the step size. This calculation is computationally demanding and therefore, is avoided in conjugate gradient algorithm by performing a line minimisation along the search direction to evaluate β_s in each step [65]. Nevertheless, a line-search is still computationally expensive as it needs the evaluation of the cost function at each line minimisation. SCG algorithm has been introduced as a way to avoid the line-search by evaluating the term Hd_s instead of using line minimisation to compute the step size β_s. Hd_s can be computed at a low computational cost by a method, introduced by LeCun et al. [100] which approximates the product of Hessian matrix with an arbitrary vector d_m without computing the full Hessian. For computing β_s using (4.16), the Hessian must be positive definite; otherwise, β_s can become negative which would lead to an increase in the cost function at each weight update accordingly. SCG algorithm tackles this problem by modifying the Hessian as:

$$\widetilde{H} = H + \alpha I \tag{4.17}$$

where H is the old value of the Hessian matrix and $\widetilde{H}$ is its modified value. I represents a unit matrix multiplied by the positive coefficient α, where α is defined such that the new Hessian $\widetilde{H}$ would be positive definite. The training is completed when the system exceeds a specified minimum amount of gradient performance or a maximum number of iterations. These two stopping criteria are employed in conjunction to avoid overfitting and to provide a good generalization performance.

During the test process, test samples with unknown outputs to the network are fed into the input layer and the network's classification outcome is computed using (4.1) and (4.2). The output would be in the range [-1 1], as the output neuron's transfer function is a hyperbolic tangent bounded to the range [-1 1]. Outputs in the range [-1 0] are interpreted as output 0 and [0 1] as output 1. The network outputs are then compared with the desired target values and the network's performance is evaluated in terms of the percentage of the test samples classified correctly which is referred to as network's accuracy.

4.2.3. Advantages

MLPs owe their extensive application in cancer prediction to several special characteristics in their theory and structure. Some of these advantages include:

- MLPs are universal approximators. Therefore, they can realize any optimal decision boundary with arbitrary precision [101, 102].
- ANNs are capable of approximating the posterior probabilities. This property makes them a suitable approach for statistical analysis [103].

Although ANNs are capable of extracting complex and nonlinear relationships from data, they have some intrinsic limitations which have made their application in medical diagnosis controvertible [20, 36]. Some of these limitations are listed and explained in the next section to be tackled later on in the next chapters.

4.2.4. Problems with the existing MLP models in cancer prediction

Numerous publications have been issued on the subject of neural network-based cancer prognosis, and yet none of them has presented a perfect solution to the problem. This mainly arises from the uncritical application of ANNs, and mainly MLPs, which has led to unpractical models for cancer prognosis. Neural network application in oncological diagnosis has been critically reviewed in statistical and neural network journals [20, 36, 74]. These publications have carried out a critical review on the problems with studies using neural network models in oncology.

Some common problems encountered with the use of MLPs in cancer prediction include overfitting (the inefficiency of the network to generalize for new patients) and uncritical estimation of the true classification error of the networks. Choosing the appropriate network structure for the in-hand problem is also another issue which has not found a unique solution so far. These problems hinder the widespread practical application of neural networks in cancer prognosis. Below is a detailed argument of the mentioned problems which is the forerunner to better understanding the main problems addressed in this thesis.

4.2.4.1. Overfitting

The flexibility of neural networks in nonlinear mapping and estimating complex functions are two of the main advantage of ANNs which has made them widespread in cancer diagnosis. However, naive use of MLPs in cancer applications and lack of accurate model evaluation methods leads to overfitting [20].

Overfitting or fitting implausible functions to the training data is one of the most common problems with the application of MLPs in cancer prediction. Overfitting is the inability of the network to generalize for unseen data which generally originates from fitting exceedingly complex functions to describe the probability of class membership of the data points at hand. In MLP neural networks, this occurs mainly due to the large ratio of the network parameters to the number of data points.

Unfortunately, there exists no standard method for measuring and quantifying the amount of overfitting. However, there are methods devised to detect overfitting and avoid it. One indicator to examine whether the network overfits the data is the ratio of the number of parameters in the network to be adjusted to the data size. In ANNs literature, 5 or 10 data points are required for each free parameter in the network as a rule of thumb [104]. In many ANN studies applying MLP in cancer prediction, this ratio is even less than 2 which implies the significant amount of overfitting and the inability of the devised networks to generalise for new patients [20]. To avoid this problem, one can employ a large dataset that has the same statistics as the population. However, increasing the size of data in medical applications is a cumbersome task which sometimes requires considerable time and funds to collect the desired data from patients. A solution to assess and reduce overfitting in MLPs is cross validation (CV) [44]. In this method, the training data are randomly divided into two parts: the *training* set and *validation* set. The training set is used for training the network and the validation set is used to estimate the validation performance. This measure is estimated by testing the network with the validation set as it is being trained with the training data. The validation performance is monitored during the training process and the training is stopped when this performance reaches its minimum value. This is shown in Figure 4.3.

In most training tasks, the validation performance is similar to the training performance at the beginning, up to a point where it takes an opposite direction and starts to increase as the training performance continues to decrease. This implies that the network has started to overfit the data from the point where the two performances take opposite directions as the network can no longer provide good performance for the validation set. This point is epoch 3 in Figure 4.3.

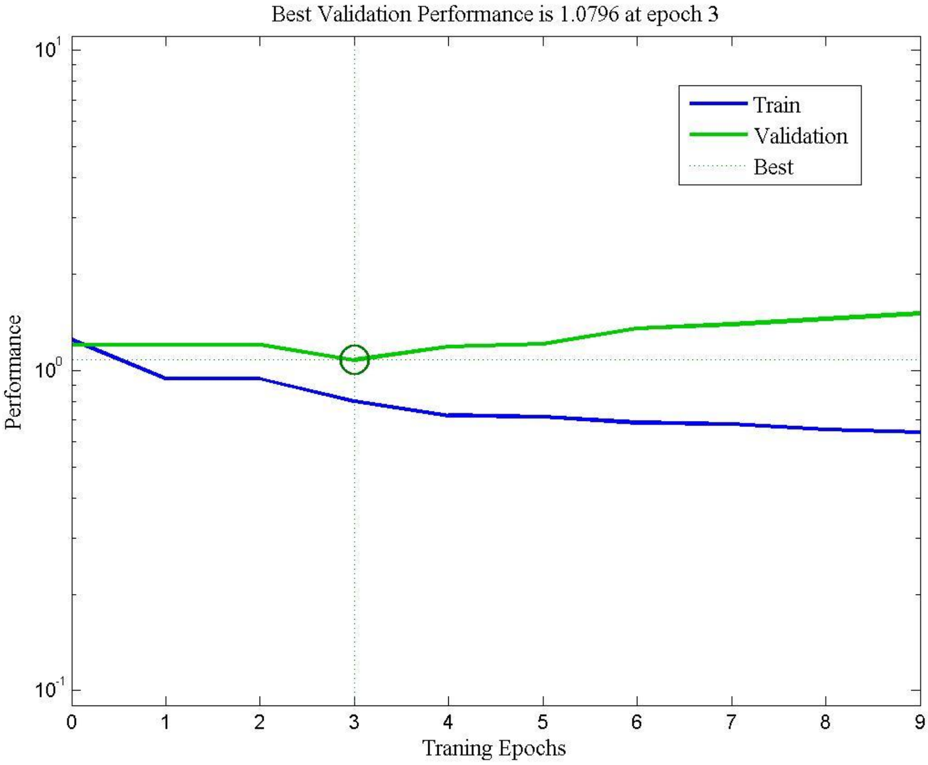

Figure 4.3: In cross validation, the training data is divided into two parts: *the training set and validation set*. The validation performance is monitored during the training process and the training is stopped when this performance reaches its minimum value (best validation performance).

CV is not an optimum method because a part of the data are kept unseen to the network as the validation set and hence, the network is trained only with a part of data. This is especially problematic with small datasets as the patterns in the data must be unveiled only with a part of the small data.

4.2.4.2. Error estimation

The realistic error of the designed network can be determined by running the network for new inputs and comparing the generated outputs with the desired targets. For this purpose, two non-overlapping groups of data are used for training and testing the network. One major issue in training and testing the ANNs employed for cancer predictions is that datasets relating to cancer can contain a relatively small number of patients in comparison with the number of measured variables. This limitation results in two problems in classification. First, the classifier will learn the existing patterns in the population using only a small number of samples. Second, an estimate of the accuracy of the classifier for new data can be only performed using the same small dataset. Using a small dataset for testing the designed network results in biased estimation of the classification error [20].

Several methods have been devised to employ the available data for training and testing the designed ANN. These error estimation methods have been discussed in section 2.7. A reliable and accurate error estimation method for evaluating the proposed ANNs in this study is fully discussed in this chapter in section 4.4.

4.3. Probabilistic neural networks

PNN is a special design of the RBF for dichotomous outputs. RBF networks typically are made up of three layers including input layer, a hidden layer with an activation function in the form of a nonlinear radial basis function and a linear output layer. The input data, provided to the network through the input layer nodes, is nonlinearly transformed into a high-dimensional hidden space by the basis functions. Afterwards, the response of the network is determined from a linear transformation of the hidden layer outputs. The PNN's overall structure is similar to that of the RBFs. Moreover, the PNN classification framework follows the same principles as the RBF that is transforming the data nonlinearly into a high-dimensional hidden space. In the following sections, some principles of the dichotomous classification in the PNN, the Bayesian probability theory and its role in developing the PNN's classification algorithm are explained in detail.

4.3.1. Probabilistic framework for classification

In a probabilistic approach to classification, features and classes are assumed as random variables for which a probability density function (PDF) can be estimated. In this context, learning from examples can be performed by estimating the PDF of the feature space and classes from the available data.

When dealing with a classification problem with two output categories, the output class can be viewed as a discrete random variable that is constructed as a binary vector O based on known class categories. Each element in O corresponds to an observation (patient) and is either 1 for patients with metastasis in the ALNs or 0 for those with no metastasis. The ultimate goal of the classification is defining a decision rule in order to determine the probability of the output class random variable being equal to 0 or 1 for a new patient with known features and unknown diagnosis.

Conventional learning from data in MLP training can be described from statistical viewpoint as maximum likelihood (ML) estimation. This happens as the network weights are adjusted during the training such that their fit to training data is maximised. In the context of probability, MLP gives a direct estimate of the posterior probability without estimating the prior or conditional probabilities. Unlike MLP, learning in PNN is based on Bayesian statistics. In Bayesian statistics, learning is based on probability distributions that are either known or can be estimated. These probability distributions describe the uncertainty being learned from the existing relationships in the data.

Bayesian statistics employs the well known Bayes theorem as a means to obtain a probabilistic model of the observed data. In this section, Bayes theorem is described briefly. Then, the PNN approach to classification using Bayes theorem is detailed.

4.3.2. Bayesian probability theory

According to Bayesian probability theory, the probability of an observation belonging to class $O = j$ ($j=1,2$ for a binary classification problem) can be expressed using Bayes theorem. This is expressed in (4.18) and (4.19) where $P(X|O = j)$

denotes the class-conditional PDF also known as likelihood function $f_j(X)$, $P(O = j)$ denotes prior probability (π_j), $P(X)$ denotes the evidence (a normalizing constant to ensure that the total probability is one and $P(O = j|X)$ denotes the posterior probability:

$$P(O = j|X) = \frac{P(X|O = j)\,P(O = j)}{P(X)} = \frac{f_j(X)\,\pi_j}{P(X)} \tag{4.18}$$

$$P(X) = \sum_{j=1,2} P(X|O = j)\,P(O = j) \tag{4.19}$$

where $X = [x_1\ x_2\ \ldots\ x_p]$ is the input vector to be classified as belonging to output class $O = j$; having two classes $j = 1{,}2$.

The probability of the output being equal to 0 or 1 for a new patient with known features and unknown diagnosis is denoted as $P(O = j|X)$ which is the posterior probability of the class O given the input X. Based on Bayes theorem in (4.18) and (4.19), posterior probability can be derived from two quantities: prior probability and the class-conditional probability density function. These two quantities can be derived by using sample data with known input features and output classes.

The prior probability $P(O = j)$ is the probability of randomly selected samples with class $O = j$ from a set of data. For a two class problem with equal probability for each class, the probability of randomly selecting a sample belonging to class 1 is 50%. Hence, the highest error given all samples are classified as 1 is 50%. The prior probability becomes noticeable in classification problems having classes with different priors. For example, in a binary classification problem with class 0 having a prior probability equal to 0.8 and class 1 with prior equal to 0.2, taking pure chance of assigning class 1 to all inputs results in 80% error. Therefore in this case, the maximum error with a dumb classifier is 80% and not 50%.

The class-conditional PDF is the conditional density of input X belonging to class $O = j$ or the likelihood of the class $O = j$ given the input X and is denoted as $P(X|O = j)$. A new observation can be classified using Bayes probability theory by assigning it to the class with the highest posterior probability. This minimises the misclassification probability and obtains the Bayes optimal decision [44].

4.3.3. Bayesian Analysis for PNN

The PNN classifies each new pattern as a member of one of the two or more output classes by modelling a Bayesian classifier. This is carried out by computing the posterior probability of a new pattern given different output classes and assigning it to the class having the largest posterior probability. As mentioned before, Bayes' theorem allows the calculation of the posterior probability by computing the multiplication of conditional probability (likelihood) and prior probability divided by a normalizing constant. The constant $P(X)$ in the denominator of (4.18) is the probability of the available input data irrespective of the knowledge about the output distribution. Hence, it can be eliminated as it is a common factor in the comparison of different posterior probabilities. Therefore, (4.18) can be simply stated as:

$$Posterior\ Probability \propto Liklihood \times Prior \qquad (4.20)$$

The prior probability is known and in the PNN with two output classes is considered as the relative frequency of training samples from each class. Thus, the key issue in computing the posterior probability is estimating the probability densities of the training patterns belonging to each class.

In the PNN, this is carried out using Parzen-Window density estimation, in which the PDF of a set of given patterns is estimated in a nonparametric manner by superimposing a set of kernels (window functions) placed on each data point [105]. Therefore, assuming two groups of data to be classified, the PDF of each group $f_j(X)$ can be estimated as:

$$f_j(X) = \frac{1}{p_j s^n} \sum_{p=1}^{P_j} \sum_{i=1}^{n} k\left(\frac{x_i - x_{ip}}{s}\right) \qquad (4.21)$$

where x_{ip} is the p^{th} training pattern and the index i refers to each element of the n-dimensional input vectors. P_j indicates the number of training patterns in each class. $k(.)$ is the kernel (window function) placed on each training pattern and its width is specified by s, denoted as “spread” in the PNN.

In the PNN, classification is carried out by estimating the $f_j(X)$ for each group. A PNN consists of four layers including input, pattern, summation and output layers [24]. The pattern layer consists of RBF nodes. Therefore, PNN is considered as a variant of RBFs.

A PNN structure with a two-class output is illustrated in Figure 4.4. For training the network, the training samples x_{ip} are presented at the input layer. During the training mode, the weights between the input and pattern layers w_{ip} are set equal to the training samples as:

$$w_{ip} = x_{ip} \tag{4.22}$$

Therefore, the size of input and pattern layers are determined by the input vector's dimension n and the number of training samples P respectively, where the number of training patterns P is the sum of the number of samples in two groups, P_1 and P_2. The size of summation layer is determined by the number of groups to be classified. Hence, in the present application the summation layer consists of two units for which the weights w_{pj} are adjusted to 1 if the training pattern belongs to the class associated with that unit and is 0 otherwise. Thus, w_{pj} is formulated as:

$$w_{pj} = \begin{cases} 1 & if \quad X \in Classj \\ 0 & if \quad X \notin Classj \end{cases} \tag{4.23}$$

Applying PNN's weight structure to (4.21) obtains:

$$f_j(X) = \frac{1}{p_j s^n} \sum_{p=1}^{P} w_{pj} \sum_{i=1}^{n} k\left(\frac{x_i - w_{ip}}{s}\right) \tag{4.24}$$

A multivariate Gaussian kernel is commonly used in a Parzen-Window PDF estimator and hence $f_j(X)$ in (4.24) can be represented as:

$$f_j(X) = \frac{1}{(2\pi)^{\frac{n}{2}} p_j s^n} \sum_{p=1}^{P} w_{pj} \sum_{i=1}^{n} exp(-\frac{\left\|x_i - w_{ip}\right\|^2}{2s^2}) \tag{4.25}$$

where *s* is the Gaussian kernel bandwidth (spread). The implications of the spread and the method for deciding an appropriate value for the spread are discussed in the section 4.3.3.1, "Spread".

Finally, a comparison is made at the output layer between the summation of kernel functions obtained for different classes, $f_j(X)$ *for j=1,2,* and the class with the highest value would be assigned to the input data. Therefore, the output decision function can be represented as:

$$O = \begin{cases} 1\ (X\epsilon\ Class1) & if \quad f_1(x) \geq f_2(x) \\ 2\ (X\epsilon\ Class2) & if \quad f_1(x) < f_2(x) \end{cases} \tag{4.26}$$

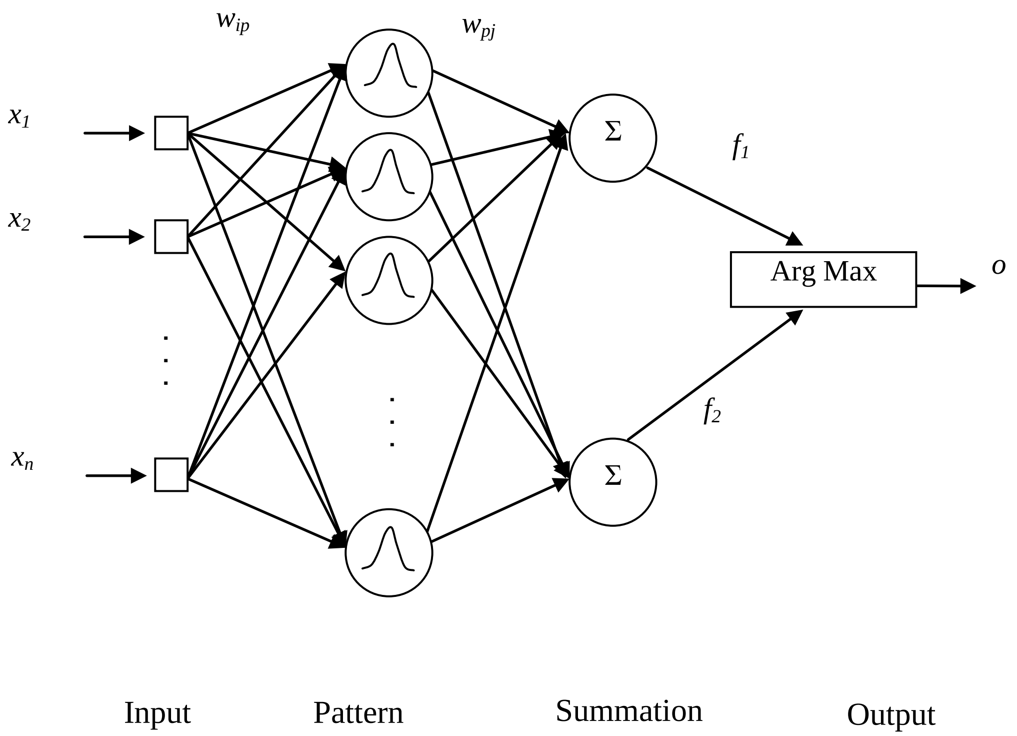

Figure 4.4: A two-class output PNN structure

4.3.3.1. Spread

Spread is the scaling factor of the window function in Parzen Window estimation which determines the width (standard deviation) of the Gaussian kernel. It is also called the smoothing factor or bandwidth by some authors. Parzen Window density estimation can be viewed as the sum of window functions centred at each observation. While the kernel functions *k* determine the shape of the windows, the spread *s* determines their width. Spread controls the appearance of the density estimate $f_j(x)$ in (4.25), which in turn signifies the amount of separation of the input patterns.

The effect of varying the spread on the shape of the $f_j(x)$ is illustrated in Figure 4.5, Figure 4.6 and Figure 4.7 for spread values of *0.5*, *1* and *2* respectively. A small spread value ($\ll$ 1) would result in probability density estimates that are too narrow with a sharp peak as illustrated in Figure 4.5 for $s = 0.5$. On the other hand, a large value for the spread gives rise to smoothed boundaries that do not separate the data effectively (Figure 4.7). The decision boundaries and the degree of generalisation capacity of the PNN depend on the choice of the spread value.

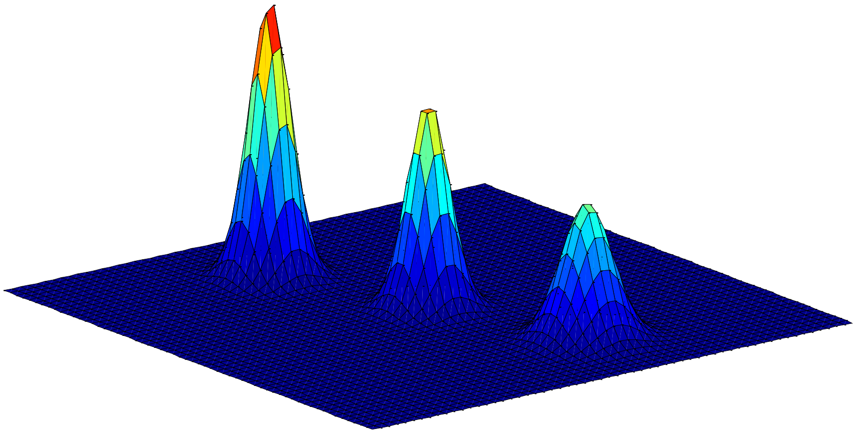

Figure 4.5: Three 2-dimentional Gaussians with s = 0.5

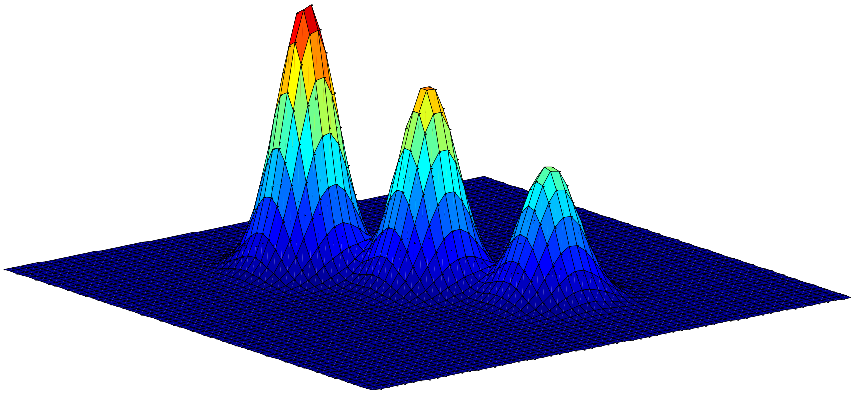

Figure 4.6: Three 2-dimentional Gaussians with s = 1

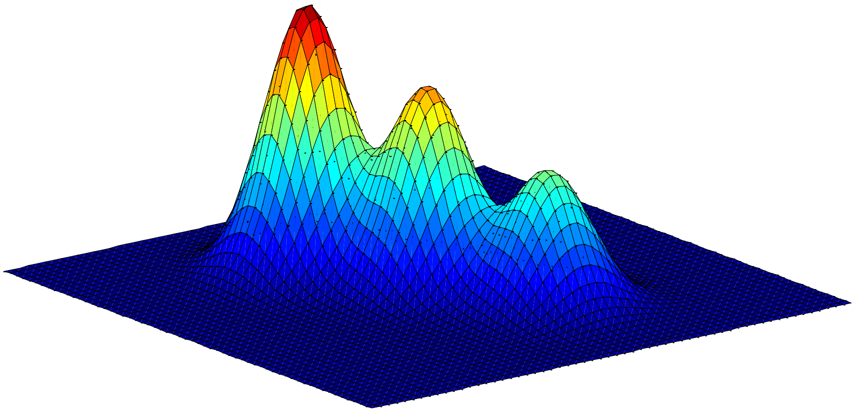

Figure 4.7: Three 2-dimentional Gaussians with s = 2

Silverman's rule-of-thumb bandwidth selection method is used for automatic derivation of spread where the spread $s = 1.06\ P^{-1/5}\ \sigma$, with σ the sample standard deviation, and P the size of training set [106]. In effect, pattern and summation layers of a PNN form a Parzen-window PDF estimator in which a set of Gaussian windows centred at each training pattern are superimposed to estimate the PDF of the training data.

4.4. The evaluation procedure

Non-realistic estimation of the misclassification probability is a common problem in most studies that have employed ANN for cancer prognosis. This has resulted in the underestimation of the prediction error rate of many proposed models [20]. This problem originates from the limited dataset which is a common issue in most medical applications. The attempt in this study has been to analyse different methods of error estimation and to quantize the amount of underestimation in various models. The ultimate goal is to introduce a standard method for defining the prediction error in neural network-based prognosis which would be compatible with limited datasets.

To achieve this aim, bootstrap .632 has been employed to evaluate the predictive accuracy of the designed PNN. In the following section, a comprehensive explanation of this technique is provided.

4.4.1. Bootstrap .632

In bootstrap method, a new dataset, called bootstrap sample, is generated by uniformly sampling P data points with replacement from the original dataset of size P. Therefore, the bootstrap sample has the same number of data points as the original data while some data points might appear in the new set more than once and some might not be included. This new data are then used to train the classifier and those data points left out from the new set are exploited to test the classifier. In the basic bootstrap, the final classifier error is estimated by replicating this procedure for B times producing D_b new datasets with $b = 1,2, \dots, B$ and averaging over all the obtained errors. B is the number of bootstrap replications. By sampling the dataset with replacement, the probability of any data point p, not being sampled after P times is:

$$prob(p \notin D_b) = (1 - \frac{1}{P})^P \cong e^{-1} \cong 0.368 \qquad (4.27)$$

Thus, the error estimation by basic bootstrap tends to be pessimistic (biased upward). The .632 bootstrap has been introduced to overcome this upward bias by computing the error using a weighted combination of the upward-biased basic bootstrap and downward-biased apparent errors [107]. This is formulated as:

$$E^{.632b} = \mathrm{A} \times E^b + (1\text{-}\mathrm{A}) \times E^a \qquad (4.28)$$

where $E^{.632}$, E^b and E^a are .632 bootstrap, basic bootstrap and apparent errors respectively and weights are defined as *A = 0.632* and *1-A = 0.368*. The upward bias of basic bootstrap and downward bias of apparent errors are mitigated in .632 bootstrap by employing a convex combination of the weights *A* and *1-A*, where convex means that the weights used to combine the two errors are nonnegative and give one when added together [108]. The .632 bootstrap has been shown to provide a reliable error estimate for neural network classifiers with a small sample size and non-zero apparent error [28]. Hence, the .632 bootstrap is employed as the error estimation method for both MLP and PNN in this study.

4.5. Choosing the best marker combination for diagnosis

Breast cancer metastasis is a complicated process. Each biomarker can only represent certain aspects of the disease. Therefore, a combination of markers is required to predict the metastasis. The biomarkers themselves have an impact on each other making the prediction task even more complex. Hence, it is important to choose a combination with markers that each represent an aspect of the disease and at the same time, do not have an overlap in the knowledge they represent about the disease. The information overlap provided by different markers complicates the prediction process and may expose the developed model to loss of generalisation. This may happen as a model with large number of input markers becomes more complex and overfitted to the presented patients with low generalisation for new ones. In addition, employing more markers for prediction is more costly to the health centres.

From the technical view point, the accuracy of the likelihood $P(X|O=j)$ estimation in (4.18) depends on the number of available samples: the larger the sample population, the closer the likelihood is to the true conditional probability of the population. This is also affected by a problem known as the *curse of dimensionality*. This means that the number of samples needed to estimate the likelihood increases exponentially with the number of available features for classification [109]. Therefore, it is important to choose only a selection of features that have a significant effect on the classification accuracy.

In addition, in the medical domain, measuring each biomarker means more cost and time to the laboratories. Therefore, obtaining the best prognosis based on the least number of features is desirable. There are myriad of feature selection methods for neural networks available in the literature. Each selection method uses a different criterion to choose the best features among several features or discard the ineffective features from the feature set. However, the most effective method in choosing the best selection of features is testing all the possible combinations of them and selecting the most effective feature or combination of features.

4.6. Summary

This chapter has presented the individual methods employed for classification of the breast cancer dataset and evaluation of the outcome. Among various ANN architectures, MLP and has been explained as the most widely used method for cancer prediction. Two learning algorithms for the MLP including the BP and the SCG have been covered in this chapter. BP is extensively implemented for training MLPs. However, it suffers from some limitations such as local minima. These limitations can be obviated partly by using conjugate gradient methods. Nevertheless, MLPs suffer from some common drawbacks including network’s complexity, numerous network variables required to be optimised and random weight initialisation.

Afterwards the PNN has been proposed as the classification method in this analysis as it offers some unique characteristics in this research. The PNN addresses some common issues in classification using limited and complex datasets available for breast cancer prediction. The PNN minimises the misclassification probability and obtains the Bayes optimal decision by computing the posterior probabilities using the model of a Bayesian classifier. The only parameter to be adjusted in the PNN is spread that has been explained in detail in section 4.3.3.1. The spread determines the degree of generalisation capacity of the PNN and hence, apt choice of the spread plays a significant role in accurate classification of both the available data and the unseen data.

To address the non-realistic estimation of the misclassification probability as a common problem in most studies using ANN for cancer prognosis, .632 bootstrap has been proposed as an accurate and reliable error estimation method. .632 bootstrap has been explained in detail in this chapter. It is a part of the resampling estimation method and an improvement over the basic bootstrap which obtains low bias and low variance especially in conjunction with small datasets that addresses the gap in accurate and reliable error estimation for small size breast cancer datasets.

In the next chapter, the results of the PNN are compared with a feed-forward MLP. In addition, to choose a reliable error estimation method for the ANNs employed in this study, the variance of accuracies obtained by .632 bootstrap, 5-fold cross validation and holdout methods are computed and compared in chapter 5.

Chapter 5

5. Experimental results and analysis

5.1. Introduction

In this chapter, the results achieved by two ANN structures - the PNN and the MLP are applied to the six markers presented in Tables 5.1 and 5.2. The first section of the chapter will describe the distinct subsets of chosen markers and how they are coded to investigate and compare the predictive importance of individual markers and their potential interaction with each other.

In the second section, the method of .632 bootstrap error estimation is compared with the performance of holdout and 5-fold CV methods for both PNN and MLP. These methods are typically chosen as the prevalent evaluation methods employed in breast cancer studies using ANN [20]. The comparison is performed by evaluating the variability of error obtained from the MLP and the PNN over 100 runs.

After choosing the best error estimation method for the PNN and the MLP, the results achieved by these two ANN structures applied to all marker combinations presented in Table 5.1 will be tabulated. Finally, the best marker combinations will be chosen.

5.2. Biomarker combinations

Distinct subsets of the six biomarkers, listed in chapter 3, are devised in the form of 1, 2, 3, 4, 5 and 6-marker combinations to present to the designed ANNs. These combinations are coded with numbers 1 to 63 for convenience. These marker combinations and their corresponding *group numbers* are displayed in Table 5.1. In this table, the first column indicates the combination group number while the markers considered in the study are listed in the first row. The following rows of the table have a value of 1 underneath for the markers already included in the combination and 0 for the markers not included. Using different marker combinations allows for better importance evaluation of different subsets for nodal prediction and also comparison of the predictive importance of individual markers and their potential interaction with each other.

Table 5.1 All possible combinations of the 6 biomarkers coded with numbers in the first row and indicated by 1 and 0 for the marker(s) included in the combination and not included respectively

Group number	Tumour size	ER	PR	p53	Ki-67	Age
1	0	0	0	0	0	1
2	0	0	0	0	1	0
3	0	0	0	1	0	0
4	0	0	1	0	0	0
5	0	1	0	0	0	0
6	1	0	0	0	0	0
7	0	0	1	0	0	1
8	0	0	0	0	1	1
9	0	0	0	1	0	1
10	0	0	0	1	1	0
11	0	0	1	0	1	0
12	0	0	1	1	0	0
13	0	1	0	0	0	1
14	0	1	0	0	1	0
15	0	1	0	1	0	0
16	0	1	1	0	0	0
17	1	0	0	0	0	1
18	1	0	0	0	1	0
19	1	0	0	1	0	0
20	1	0	1	0	0	0
21	1	1	0	0	0	0
22	0	0	0	1	1	1
23	0	0	1	0	1	1
24	0	0	1	1	0	1
25	0	0	1	1	1	0
26	0	1	0	0	1	1
27	0	1	0	1	0	1
28	0	1	0	1	1	0
29	0	1	1	0	0	1

Table 5.1 continued

Group number	Tumor size	ER	PR	p53	Ki-67	Age
30	0	1	1	0	1	0
31	0	1	1	1	0	0
32	1	0	0	0	1	1
33	1	0	0	1	0	1
34	1	0	0	1	1	0
35	1	0	1	0	0	1
36	1	0	1	0	1	0
37	1	0	1	1	0	0
38	1	1	0	0	0	1
39	1	1	0	0	1	0
40	1	1	0	1	0	0
41	1	1	1	0	0	0
42	0	0	1	1	1	1
43	0	1	0	1	1	1
44	0	1	1	0	1	1
45	0	1	1	1	0	1
46	0	1	1	1	1	0
47	1	0	1	0	1	1
48	1	0	1	1	0	1
49	1	0	1	1	1	0
50	1	1	0	1	0	1
51	1	1	0	1	1	0
52	1	1	1	0	0	1
53	1	1	1	0	1	0
54	1	1	1	1	0	0
55	1	1	0	0	1	1
56	1	0	0	1	1	1
57	0	1	1	1	1	1
58	1	0	1	1	1	1
59	1	1	0	1	1	1

Table 5.1 continued

Group number	Tumour size	ER	PR	p53	Ki-67	Age
60	1	1	1	0	1	1
61	1	1	1	1	0	1
62	1	1	1	1	1	0
63	1	1	1	1	1	1

The ‘Group number’ in the above table will be frequently referenced throughout this and the next chapter to reference each marker combination.

5.3. Choosing a reliable network evaluation method

In this section, three network evaluation methods for the designed ANNs are compared. Evaluating the classifier performance for DSSs is normally carried out by estimating the area under curve (AUC) of the receiver operating characteristic (ROC) curve. The ROC is a plot of classifiers’ true positive rate (TPR) against its false positive rate (FPR). The TPR is defined as the *sensitivity* of the classifier while the FPR is defined as $1 - specificity$. Sensitivity and specificity will be described in the next section as the performance indicator for the designed ANNs. In this section, the focus is on estimating the classifier’s AUC with a limited sample size. This has been previously addressed for linear classifiers [28, 110]. For ANN classifiers, the problem of AUC estimation with limited data size has been tackled by Sahiner et al. [28, 111].

To obtain an estimate for the ANN's AUC, the network is trained with a part of the available data and tested with the remaining data, kept unseen to the network. The predicted AUC ($\widehat{AUC}$) is therefore an estimate of the true AUC when the network trained with N samples tested with the true population. To assess the reliability of the $\widehat{AUC}$, the training and test experiments should be repeated for N times. The mean and standard deviation of the predicted AUC are then computed to provide a measure for comparing the reliability of different evaluation methods and estimating the network's performance:

$$\mu\left(\widehat{AUC}_m\right) = \frac{1}{N}\sum_{n=1}^{N}\widehat{AUC}_{n,m} \tag{5.1}$$

$$S^2\left(\widehat{AUC}_m\right) = \sqrt{\frac{1}{N-1}\sum_{n=1}^{N}[\widehat{AUC}_{n,m} - \mu\left(\widehat{AUC}_m\right)]^2} \tag{5.2}$$

where $\widehat{AUC}_{n,m}$ is the predicted AUC for the network evaluated with m^{th} method in the n^{th} experiment. Running the experiment for N times is similar to a Monte Carlo simulation in that the process is replicated several times and the mean and standard deviation of the results are computed and assessed as a reliability measure. In Monte Carlo method, the experiments are repeated each time by drawing a sample from the true population. This is possible only when a there is access to the true population or its distribution so that different samples would be available to carry out the experiments. However, in this thesis, the sample size is limited and the data distribution is unknown. Therefore, each experiment should be run by applying different training and test data that are obtained by sampling from the available dataset. The number of times that the experiments should be carried out are set to N=100. In the following sections, the approach for obtaining the AUC for the three evaluation methods evaluated in this research will be explained.

5.3.1. Holdout

For holdout estimate of the ANN performance with a dataset of size P, the data should be divided into two parts called training data P_{train} and test data P_{test}=P- P_{train}. The process of dividing data is carried out for N=100 times and the average and standard deviation of the AUC are obtained.

5.3.2. K-fold Cross validation

Another method that is considered for estimating the ANN performance is K-fold CV. To predict AUC in this method, the data are partitioned into K parts. Each part is then chosen as the test data while the network is being trained with the rest of the data. The final AUC is predicted by averaging over the results obtained for each part:

$$\widehat{AUC}_{CV} = \frac{1}{K}\sum_{k=1}^{K}\widehat{AUC}_{k,CV} \qquad (5.3)$$

Similar to holdout method, the process of data partitioning is repeated for N=100 times and the average and standard deviation of the AUC is obtained.

5.3.3. .632 Bootstrap

Considering $\mathbf{x} = \{x_1 x_2 \dots x_P\}$ as the available data of size P, where the bold letter **x** denotes a set of data and italic letters x_i represent a data vector (i.e. a patient information of different biomarkers and age), the true distribution of data can be considered as F. In bootstrap, an empirical distribution of $\hat{F}$ is considered such that each x_i has a probability of $1/P$. A bootstrap sample indicated by $\mathbf{x}^* = \{x_1{}^* x_2{}^* \dots x_P{}^*\}$ with size P can be created by sampling with replacement from **x** or in other words, randomly sampling from the empirical distribution $\hat{F}$. As mentioned in section 4.4.1, the classifier performance obtained from the bootstrap samples is biased. In order to estimate the true performance of a classifier in bootstrap method, this bias should be computed and subtracted from the bootstrap performance [112].

Since the reliability of different error estimation methods is being measured by AUC, a AUC (S_{train}, S_{test}) is chosen to represent the AUC of a classifier when trained with S_{train} and tested with S_{test}. Considering B bootstrap datasets represented as $\mathrm{x}^{*1}\mathrm{x}^{*2} \dots \mathrm{x}^{*B}$, each set is of size P, obtained by randomly sampling with replacement from the original dataset x, can be represented as $\mathrm{x}^{*b} = \{x_1{}^{*b} x_2{}^{*b} \dots x_P{}^{*b}\}$. With this description, there exists a subset of dataset x which is not included in b^{th} bootstrap sample x^{*b}, indicated here as $\overline{\mathrm{x}^{*b}}$. The performance of the classifier can then be obtained by training with bootstrap sample x^{*b} and testing with the remaining samples $\overline{\mathrm{x}^{*b}}$, i.e. $AUC(\mathrm{x}^{*b}, \overline{\mathrm{x}^{*b}})$. The final AUC, obtained from B bootstrap samples is computed by averaging over all the computed bootstrap results as:

$$AUC_b = \frac{1}{B}\sum_{b=1}^{B} AUC(\mathrm{x}^{*\mathrm{b}}, \overline{\mathrm{x}^{*\mathrm{b}}}) \qquad (5.4)$$

where AUC_b is the bootstrap AUC. As discussed in section 4.4.1 [112], the classifier performance obtained by this method is biased upward (pessimistic). To remove this upward bias, the .632 bootstrap AUC is computed as the combination of bootstrap AUC and the apparent AUC. The apparent AUC can be achieved by training and testing the classifier with all available data as:

$$AUC_a = AUC(\mathrm{x}, \mathrm{x}) \qquad (5.5)$$

From equations (4.10) and (4.11), the .632 bootstrap AUC is estimated by combining the optimistic apparent AUC and pessimistic bootstrap AUC as:

$$\widehat{AUC}_{.632} = .368.AUC(\mathrm{x}, \mathrm{x}) + .632.AUC_b \qquad (5.6)$$

5.3.4. Evaluation methods for the PNN

To investigate the reliability of .632 bootstrap in estimating the output error of the PNN, it is compared with the two commonly employed error estimation methods in ANNs - 5-fold CV and holdout methods. For this purpose, the network is run for 100 times for each of the 63 possible marker combinations and the estimated AUC by .632 bootstrap, 5-fold CV and holdout methods are computed. Table 5.1 shows these 63 possible marker combinations including subsets of samples with 1, 2, 3, 4, 5 and

6 dimensions. For each of these dimensions, the three error estimation methods are compared in Figure 5.1 in the form of boxplots of the computed *AUC*s over 100 runs. The x-axis in Figure 5.1 shows the input dimension (number of markers included in the combination) and the y-axis is the computed *AUC*. For .632 bootstrap error estimation in the PNN, $B = 50$ bootstrap replications is used as this number is proved to be sufficient for most applications [75].

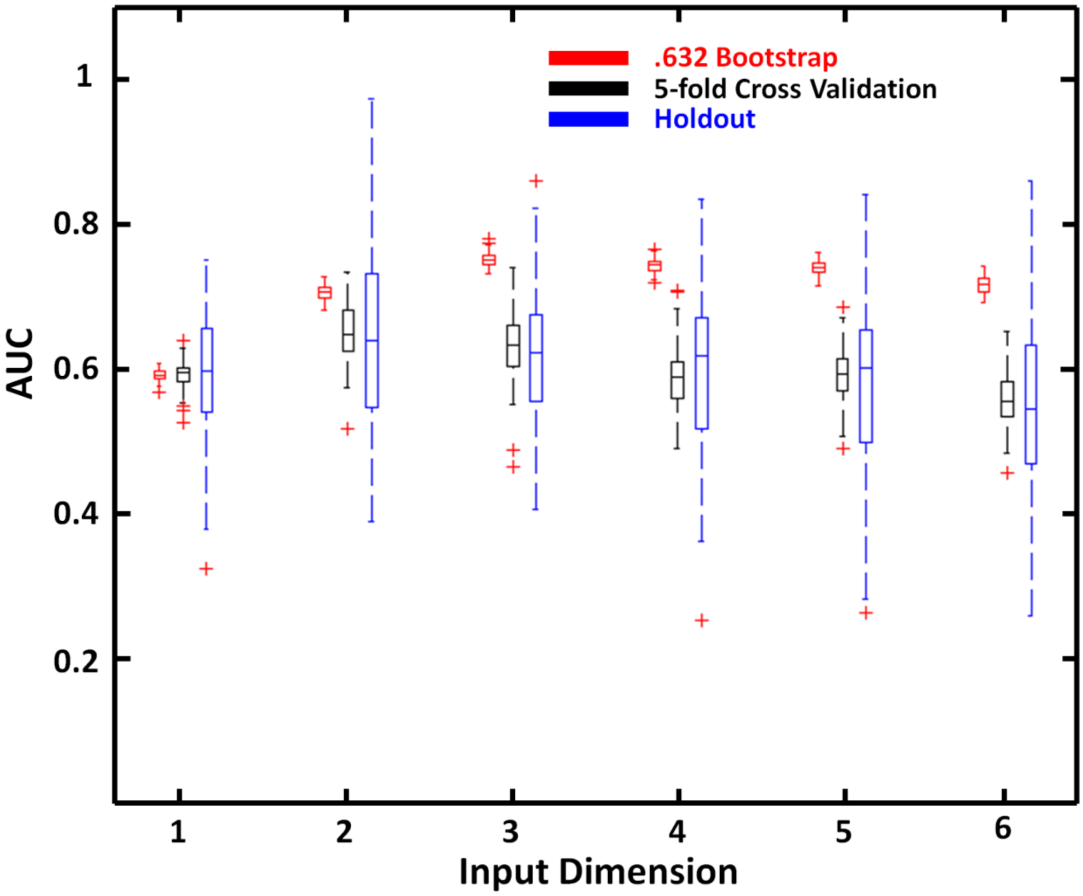

Figure 5.1: The comparison of different error estimation methods for the PNN. These methods include .632 bootstrap, 5-fold cross validation and holdout methods for which the variability of the *AUC* is illustrated by boxplots. The Input Dimension on the x-axis shows the number of markers used as the input of the PNN.

In the above figure, the horizontal line in each box shows the median of the results for each error estimation method while the length of the lines show the degree of dispersion of the results. The '+' signs out of the boxes show the outlier values. The boxplots in this figure illustrate the high variability and existence of outliers when computing the AUC using holdout method. This variability is reduced when using the 5-fold CV and is the least when using .632 bootstrap. Regarding the medians of the results illustrated in Figure 5.1, it is clear that the median for .632 bootstrap is larger than that of 5-fold CV and holdout methods. Even so, the medians of all of the compared error estimation methods for the 3, 4, 5 and 6-marker combinations are further apart than those of the 1 and 2-marker groups.

In addition, the AUC variability is illustrated for all combinations in Figure 5.2. In this figure, the x-axis shows the group numbers which refers to the group numbers defined for each marker combination in Table 5.1. The length of each line shows the variability of the AUC while the median is demonstrated with circles. Each sub-figure in Figure 5.2 includes only combinations with the same number of markers. In this figure, the circles show the median of the results while the lengths of the lines represents their dispersion.

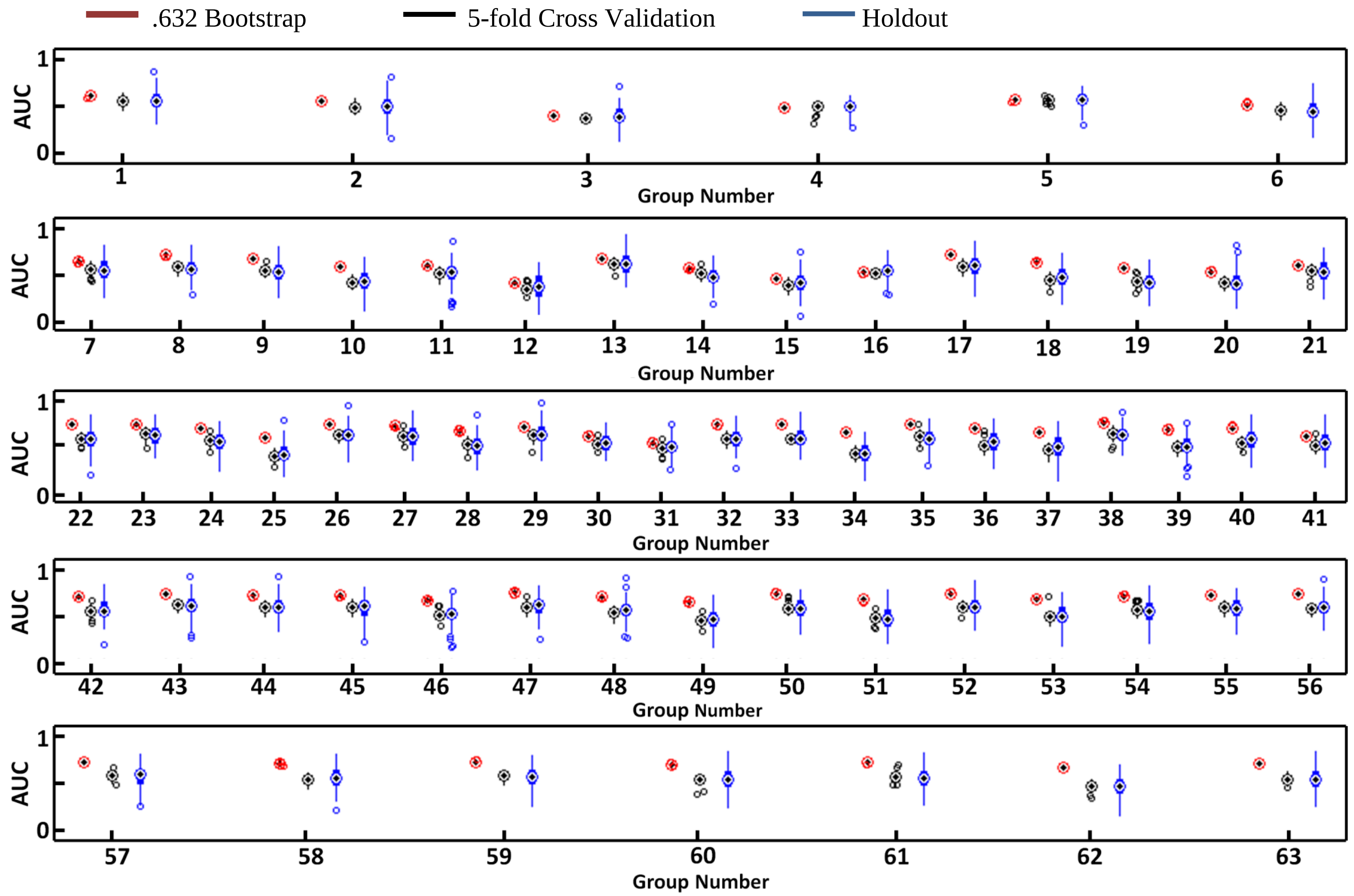

Figure 5.2: The comparison of .632 bootstrap, 5-fold cross validation and holdout methods error estimation methods for the PNN. The Group Number on the x-axis refers to the group numbers defined in Table 5.1.

This figure illustrates that the AUC variability is high for holdout method for all the combinations of the same dimension and is significantly low for the .632 bootstrap method. The difference in the median of different error estimation methods shows that .632 bootstrap achieves higher median compared to the other two in most combinations. Examining each subgroup in Figure 5.2 reveals that for the first two subgroups including the 1 and 2-marker combination groups respectively, the medians of results for all error estimation methods are close to each other. For the last three subgroups in this figure showing the 3, 4, 5 and 6-maerker combination groups, the medians are further apart. Nevertheless, the median for the 5-fold CV and holdout methods are close for all marker groups.

5.3.5. Evaluation methods for the MLP

The reliability of .632 bootstrap in estimating the output error of the MLP is investigated by comparing .632 bootstrap with 5-fold CV and holdout methods. Similar to the comparison employed in the PNN, the variability of MLP AUC over 100 runs, estimated by these three error estimation methods, is computed for each of the 63 possible combinations. To perform .632 bootstrap error estimation, $B = 50$ bootstrap replications is used.

The comparison results for combinations with different input dimensions are illustrated in Figure 5.3. These results are showed in the form of boxplots to clearly demonstrate the difference between the variability of results obtained by different methods. The x-axis in Figure 5.3 labelled as the “Input Dimension” refers to the number of markers in each group while the y-axis shows the AUC.

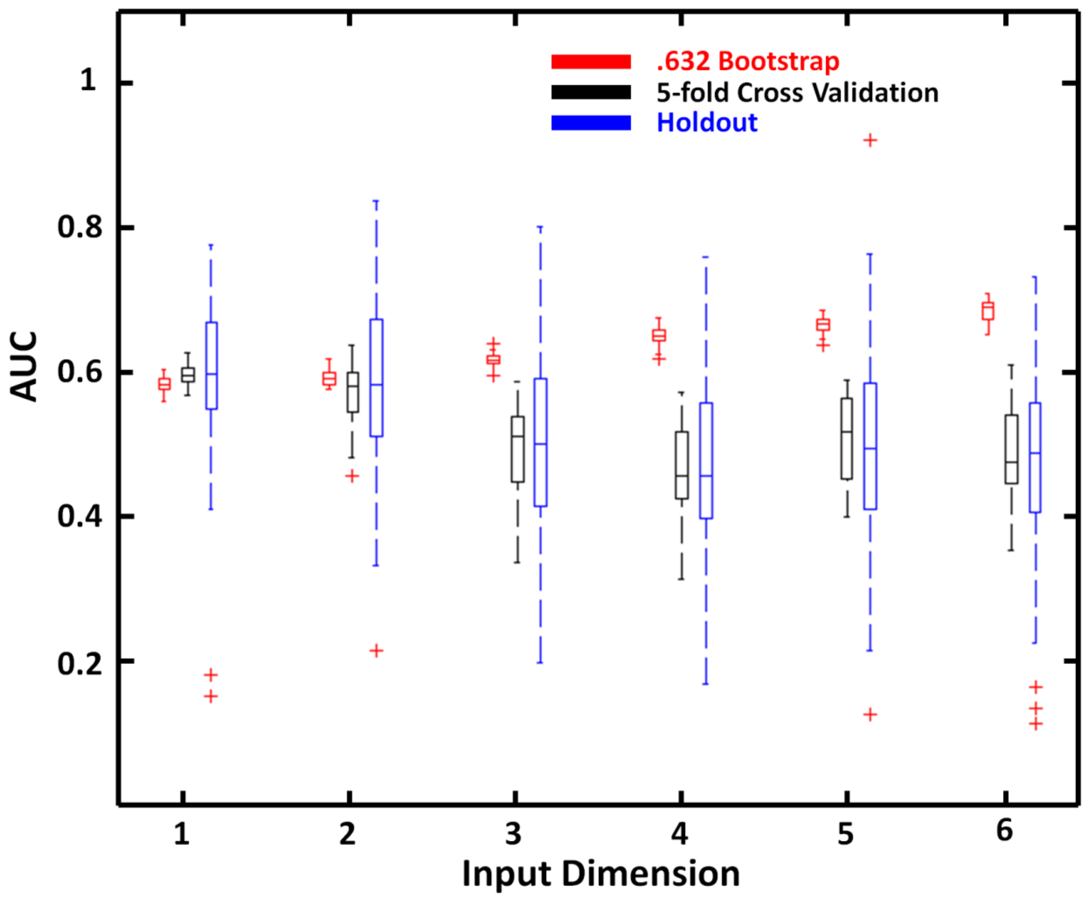

Figure 5.3: The comparison of different error estimation methods for the MLP. These methods include .632 bootstrap, 5-fold cross validation and holdout methods for which the variability of the *AUC* is illustrated by boxplots. The Input Dimension on the x-axis shows the number of markers used as the input of the PNN.

There are similar trends in Figure 5.3 and Figure 5.1 for the variability and the median of the results for different error estimation methods over different marker groups.

As it can be seen in Figure 5.1 and Figure 5.3, there are similar trends in PNN and MLP for the variability and the median of the results for different error estimation methods over different marker groups. Nevertheless, comparing these two figures reveal that the results for the MLP in Figure 5.3 maintains higher number of outliers compared to those provided by the PNN in Figure 5.1. In Figure 5.3, the variability of the results obtained by holdout method is significantly higher than that of the 5-fold CV and bootstrap. Among the latest two, .632 bootstrap obtains the lowest result variability. The median of the results for .632 bootstrap is slightly lower than that of the 5-fold CV and holdout methods for the 1-marker combination group. This trend is reversed for the rest of the input dimensions while the difference of median between .632 bootstrap and 5-fold CV and holdout methods is further apart for larger input dimensions.

Results in Table 5.4 illustrate the AUC variability obtained by the MLP evaluated with the .632 bootstrap, 5-fold CV and holdout methods for all combinations. The x-axis shows the group numbers defined for each marker combination in Table 5.1. The length of each line shows the variability of the AUC while the median is demonstrated with circles. Each sub-figure in Table 5.4 only includes combinations with the same number of markers.

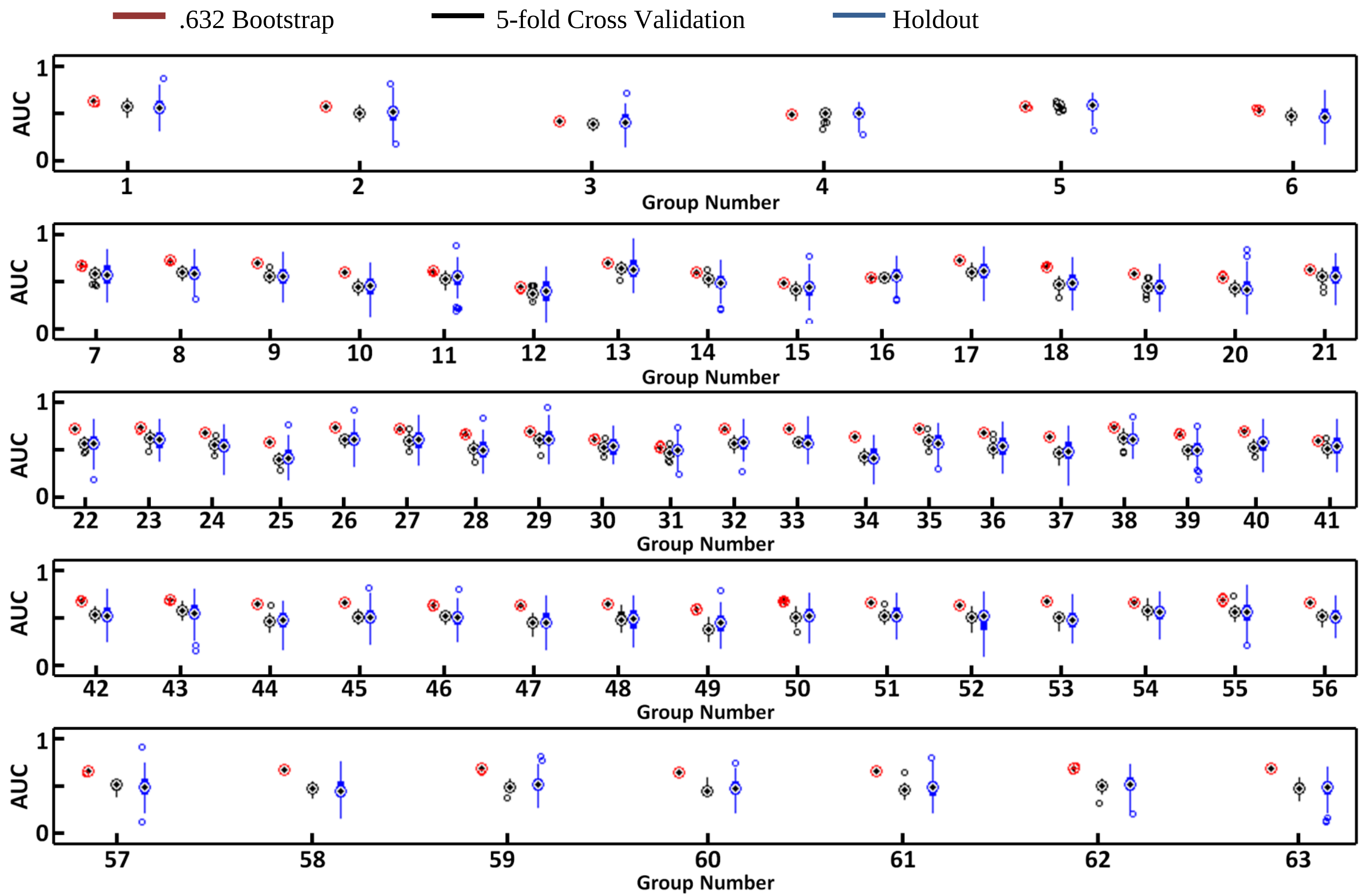

Figure 5.4: The comparison of .632 bootstrap, 5-fold cross validation and holdout methods error estimation methods for the MLP. The Group Number on the x-axis refers to the group numbers defined in Table 5.1.

This figure illustrates that the AUC variability is high for holdout method for all the combinations of the same dimension and is significantly low for the .632 bootstrap method. The difference in the median of different error estimation methods shows that .632 bootstrap method can achieve higher median compared to 5-fold CV and holdout method in most combinations. Examining each subgroup in Figure 5.2 reveals that for the first two subgroups that include the 1 and 2-marker combination groups respectively, the medians of results for all error estimation methods are close to each other. For the last three subgroups (3, 4, 5 and 6-maerker combination groups), the medians are further apart. Nevertheless, the median for the 5-fold CV and holdout methods are close for all marker groups.

5.4. Neural network evaluation for breast cancer prediction

In this section, the predictive ability of the designed networks is evaluated using sensitivity, specificity and accuracy. Sensitivity represents the ratio of the correctly diagnosed patients with metastasis to the total number of patients diagnosed with metastasis:

$$Sensitivity\ (\%) = \frac{TP}{TP + FN} \qquad (5.7)$$

This means that the denominator is obtained by summing the number of both correctly and incorrectly diagnosed patients. Specificity denotes the same ratio for the correctly diagnosed patients without metastasis:

$$Specificity\ (\%) = \frac{TN}{TN + FP} \tag{5.8}$$

Accuracy is defined as the ratio of correctly classified cases to the total number of classifications, computed as:

$$Accuracy\ (\%) = \frac{TP + TN}{TP + FP + TN + FN}, \tag{5.9}$$

where TN, TP, FP and FN percentages respectively stand for:

- True negative: the percentage of normal samples classified as normal
- True positive: the percentage of cancer samples classified as cancerous
- False positive: the percentage of normal samples classified as cancerous
- False negative: the percentage of cancer samples classified as normal

Sensitivity and specificity values reported in this study are rounded to the nearest integer and expressed in percentages. While it is desirable to have both sensitivity and specificity values at maximum of 100%, the associated risk of each of these parameters is different. A high sensitivity rate is more desirable than a high specificity rate. This is because the sensitivity indicates the number of correctly diagnosed metastasis cases which need auxiliary dissection and hence, a lower rate of mortality. On the other hand, the specificity is an indication of correctly classified normal patients who do not need a dissection and hence, a lower morbidity rate by reducing the number of unnecessary dissections. While keeping both mortality and morbidity rates at a low level is valuable, it is obvious that lower mortality is more advantageous to patients than lower morbidity.

Apparent error is also computed for each network. It indicates how well the network can separate the training data. Low apparent error indicates the efficiency of the network in classifying different categories in training data. However, low apparent error does not always come with low test error, since low apparent error can also imply that the network is over-trained. This means that the network has learnt some specific patterns in the training set which is not common to all patterns in general and eventually results in a small apparent error but a large test error.

5.4.1. PNN results

The prediction results obtained by the PNN for classifying node positive and negative tumours using all different combinations of input biomarkers are tabulated in Table 5.2. The group number in the first column of this table refers to the group numbers for each marker combination defined in Table 5.1. This table demonstrates the sensitivity, specificity, accuracy and training error computed for each group of markers using the designed PNN.

The inputs of the PNN are normalised to the range [0 1]. This normalisation is only performed for the variable vectors age and tumour size since p53 and Ki-67 are already within this range and ER and PR are discrete variables. The normalisation is carried out by subtracting the vector minimum from the vector and dividing it by its range (i.e. the difference between the vectors' maximum and minimum). This normalisation is essential to the accurate performance of the PNN as the input variables are combined in the PNN via a distance kernel function. With such functions, the range of the variables affects their influence on the outcome. Hence, if one input is within the range of 0 to 1 and another input has a larger range such as .5 to 5.5, the contribution of the second input to the distance overshadows the first input. Therefore, it is important to normalise the input variables such that their variability is a reflection of their significance. Because this is an exploratory study in which the aim is to identify the importance of each variable in determining the outcome, the inputs are normalised to the same range.

Table 5.2: Nodal predictive accuracy of entire marker set and its different subsets assessed by PNN

Group Number	Sensitivity (%)	Specificity (%)	Test Accuracy (%)	Apparent Error (%)
1	51	52	56	37
2	52	52	57	35
3	57	64	61	34
4	74	66	51	47
5	54	53	55	42
6	53	65	59	37
7	50	52	56	34
8	62	52	64	23
9	52	56	63	21
10	54	52	58	26
11	54	51	55	34
12	64	64	56	35
13	59	55	61	31
14	52	53	53	39
15	51	63	61	32
16	56	57	54	42
17	67	58	66	19
18	52	53	58	26
19	51	64	64	23
20	59	63	58	34
21	52	65	62	32
22	62	52	67	11
23	55	52	63	18
24	51	52	61	23
25	59	55	59	24
26	70	50	67	13
27	63	51	64	21
28	52	55	60	27
29	54	51	60	27

Table 5.2 continued

Group Number	Sensitivity (%)	Specificity (%)	Test Accuracy (%)	Apparent Error (%)
30	52	61	53	34
31	50	58	61	29
32	60	53	65	13
33	60	58	68	11
34	54	52	61	18
35	54	50	63	18
36	55	50	58	24
37	58	66	63	24
38	52	58	67	15
39	55	55	62	19
40	50	66	67	18
41	55	64	62	27
42	65	53	67	11
43	63	53	69	11
44	53	52	64	13
45	57	50	64	13
46	52	51	61	23
47	55	54	64	11
48	52	50	65	11
49	54	54	61	15
50	60	55	67	13
51	57	58	64	13
52	65	55	66	11
53	52	50	62	15
54	61	62	68	16
55	64	53	69	8
56	55	51	68	8
57	52	56	64	8
58	55	51	67	5
59	53	52	67	5

Table 5.2 continued

Group Number	Sensitivity (%)	Specificity (%)	Test Accuracy (%)	Apparent Error (%)
60	53	52	64	8
61	66	57	67	8
62	52	65	67	10
63	59	52	67	5

Results in Table 5.2 show that among the individual markers in groups 1 to 6, group 3 representing the p53 obtains the best prediction accuracy of 61%. Among the 2-marker combinations in groups 7 to 21, tumour size and age in group 17 achieve the highest accuracy of 66%. In addition, two more combinations including group 8 with Ki-67 and age and group 19 with tumour size and p53 also gain high accuracies close to group 17. The existence of tumour size in these combinations displays the prediction potential of this marker in combination with other markers. Moreover, p53 and tumour size obtain a high accuracy among the 2-marker combinations which proves the potential of p53 for prediction both as an individual marker and in combination with other markers.

Examining the 3-marker combinations in groups 22 to 41, group 33 with tumour size, p53 and age gains the highest accuracy of 68%. This result is very interesting as this group is a combination of the markers in groups 17 and 19 which achieved the highest accuracies in 2-marker combination groups. This can be explained by the interaction between the markers that can improve the accuracy when combined together.

Amongst groups 42 to 56 which include 4-marker combinations, the highest accuracy is achieved by two combinations in groups 43 and 55 with an accuracy of

69%. These combinations have Ki-67 and age as common biomarkers. These two markers also give a high accuracy amongst 2-marker combinations. This clearly proves the potential of this combination of markers. It is noticeable that PR is missing from all the combinations with high accuracies. This can be explained by high correlation of ER and PR ($PCC = 0.615$) which may count for the compensation of PR by using ER.

Groups 57 to 62 which constitute 5-marker combination groups shows the same accuracy except for the groups where either tumour size or p53 are taken out. This emphasizes the significant role of these markers in interaction with other markers to improve the prediction results.

There exists a trend between the apparent error of different marker combinations and their accuracies. From Table 5.2, it can be observed that by adding more markers to the combinations, the accuracy increases while the apparent error decreases. This trend is disrupted in the 5 and 6-marker combinations where the apparent error decreases substantially to as low as 5% while the accuracy does not improve any longer and even decreases slightly compared to 4-marker combinations. This may be explained by network overfitting. Using more markers at the network input means more nodes and connections which may result in overfitting the network for the training data and loss of generalisation for new data.

5.4.2. MLP results

The MLP structure designed in the present study consists of three layers: input, one hidden and an output layer. The number of nodes in the input is varied for each group of markers depending on the number of markers included as input. The number of nodes in the output layer is equal to one for all groups as it only depends on the number of groups to be classified. The number of hidden nodes is a free parameter which can be defined by trial and error considering the number of training patterns together with the input patterns' dimension. Number of hidden nodes is chosen as twice the number of input units in accordance with the minimum of four nodes in the hidden layer. This is done by calling the INITNW function in MATLAB [113] which initialises weights using Nguyen-Widrow layer initialisation function [114].

The prediction performance of the network is evaluated using the .632 bootstrap. The MLP results for all marker combinations are summarized in Table 5.3. This table demonstrates the sensitivity, specificity, accuracy and training error computed for each group of markers using the designed MLP. To avoid overfitting, the training should be stopped when it either reaches the maximum number of 1000 epochs or its performance gradient drops below a minimum value of 10^{-6}.

Table 5.3 Nodal predictive accuracy of entire marker set and its different subsets assessed by MLP

Group Number	Sensitivity (%)	Specificity (%)	Test Accuracy (%)	Apparent Error (%)
1	51	55	52	44
2	51	55	52	47
3	54	63	62	34
4	56	59	58	47
5	56	59	58	42
6	56	59	58	37
7	56	59	58	37
8	56	59	58	29
9	52	57	56	35
10	51	57	56	34
11	52	57	56	35
12	52	57	56	34
13	52	57	56	31
14	51	57	56	39
15	51	57	56	32
16	53	57	55	42
17	54	59	57	34
18	54	59	57	35
19	50	58	59	32
20	50	58	59	35
21	55	58	61	31
22	56	52	64	21
23	56	52	64	29
24	56	52	64	23
25	56	52	64	29
26	54	52	66	18
27	54	59	62	24
28	54	52	58	26
29	54	52	61	31

Table 5.3 continued

Group Number	Sensitivity (%)	Specificity (%)	Test Accuracy (%)	Apparent Error (%)
30	54	52	59	34
31	58	60	63	29
32	51	51	65	24
33	54	51	68	16
34	54	51	61	27
35	52	52	58	34
36	52	52	56	29
37	52	52	63	21
38	52	52	68	32
39	52	52	58	29
40	52	52	68	26
41	54	59	62	27
42	54	54	68	13
43	56	51	66	11
44	56	51	61	15
45	55	50	64	11
46	55	50	59	24
47	55	50	62	11
48	52	56	58	8
49	52	56	61	13
50	52	56	69	8
51	52	56	57	11
52	51	53	66	10
53	50	55	64	34
54	52	53	63	19
55	50	53	68	5
56	54	52	65	11
57	52	52	60	8
58	52	55	65	6
59	51	53	66	3

Table 5.3 continued

Group Number	Sensitivity (%)	Specificity (%)	Test Accuracy (%)	Apparent Error (%)
60	51	56	64	3
61	51	54	63	10
62	50	50	67	5
63	53	51	68	2

Predicting the nodal involvement using different combination of markers for the MLP manifests almost the same trends and relationships between markers as the PNN. From Table 5.3, p53 achieves the highest accuracy among individual markers. Similar to the PNN results, the highest accuracy in MLP results is 69%. Nevertheless, the combination that achieves this accuracy is different from the PNN. This combination includes tumour size, ER, p53 and age while the combination in the PNN with highest accuracy include tumour size, ER, Ki-67 and age. This implies that the PNN and the MLP may discover different aspects of relationships between markers due to their different approaches to pattern classification. Overall, the PNN obtains higher accuracy in most combinations.

5.5. Choosing the efficient marker combination

For comparing the PNN and the MLP results, 19 marker groups that yield the highest accuracy among all marker groups for both the PNN and the MLP are selected. The results for these groups are tabulated in Table 5.4, where the first column displays the marker groups referring to the marker groups defined in Table 5.1. All 19 selected combinations in Table 5.4 obtain accuracies above 66% for the PNN and above 56% for the MLP. The relation of markers in these best marker combinations will be further investigated and discussed in the next chapter.

Table 5.4 The comparison of nodal predictive accuracy of selected marker groups assessed by the PNN and the MLP

Group Number	Method	Sensitivity (%)	Specificity (%)	Test Accuracy (%)	Apparent Error (%)
17	PNN	67	58	66	19
	MLP	54	59	57	34
22	PNN	62	52	67	11
	MLP	56	52	64	21
26	PNN	70	50	67	13
	MLP	54	52	66	18
32	PNN	60	53	65	13
	MLP	51	51	65	24
33	PNN	60	58	68	11
	MLP	54	51	68	16
38	PNN	52	58	67	15
	MLP	52	52	68	32

Table 5.4 continued

Group Number	Method	Sensitivity (%)	Specificity (%)	Test Accuracy (%)	Apparent Error (%)
40	PNN	50	66	67	18
	MLP	52	52	68	26
42	PNN	65	53	67	11
	MLP	54	54	68	13
43	PNN	63	53	69	11
	MLP	56	51	66	11
50	PNN	60	55	67	13
	MLP	52	56	69	8
52	PNN	65	55	66	11
	MLP	51	53	66	10
54	PNN	61	62	68	16
	MLP	52	53	63	19
55	PNN	64	53	69	8
	MLP	50	53	68	5
56	PNN	55	51	68	3
	MLP	54	52	65	11
58	PNN	55	51	67	5
	MLP	52	55	65	6
59	PNN	53	52	67	5
	MLP	51	53	66	3
61	PNN	66	57	67	8
	MLP	51	54	63	10
62	PNN	52	65	67	10
	MLP	50	50	67	5
63	PNN	59	52	67	5
	MLP	53	51	68	2

Comparing the sensitivity and the specificity results obtained by the PNN and the MLP, it is evident that the PNN can achieve better prediction for both cases. This difference is more noticeable in the sensitivity results which establish the PNN as a good choice compared the MLP when a high sensitivity is the target.

5.6. Summary

The experimental results have been presented in this chapter for the three main sections of the study: 1) investigating the reliability of different error estimation methods, 2) the comparison of the PNN classification results against the MLP and 3) choosing the best marker combinations.

Section 5.3 has covered the experiments planned to investigate the reliability of different error estimation methods for the designed ANNs in this study. The results prove the low variability of the .632 bootstrap in comparison with the CV and holdout methods. This low variability guarantees the reliability of the network evaluation obtained by .632 bootstrap. The high variability in holdout and CV methods implies that reporting only the best results could be considerably misleading in cancer studies. This renders the reliability of the studies employing the holdout and CV methods for ANN's error estimation debatable.

The results of the PNN and the MLP including their accuracy, sensitivity, specificity and apparent error have been computed and compared in section 5.4. These results confirm the superiority of the PNN in obtaining more accurate results. Moreover, the sensitivity and specificity of the PNN are higher or close to that of the MLP. Nevertheless, the apparent errors obtained from the PNN are only slightly different from the MLP. This can prove that the degree of generalisation of the PNN is comparable with that of the MLP.

Section 5.5 has presented the best results obtained by the network in the same table to provide a better understating and analysis of the two ANN results by adding the best marker combinations. These results will be the subject of further investigation and discussion in the next chapter which is dedicated to the comprehensive discussion, detailed explanation and further analysis of the results.

Chapter 6

6. Discussion and analysis

6.1. Introduction

The main steps in design of a pattern classifier system are: feature selection, classifier design and classification error estimation. All these steps should be designed considering the size, dimensionality and complexity of the available data in order to achieve an optimum pattern classification system which yields maximum accuracy.

In this study, the aim was to classify breast cancer data and to estimate classification accuracy of the designed classifier. Feature selection has not been a direct aim of the study as it considered all feature combinations to investigate each feature potential in output classification. Therefore, the research focused on two stages of pattern classification system design; namely classifier design and classification error estimation considering the limitations of the available dataset – small sample size and high complexity. The stages involved in designing a pattern classifier are shown in Figure 6.1, where the stages that this is focusing on are shown in grey boxes.

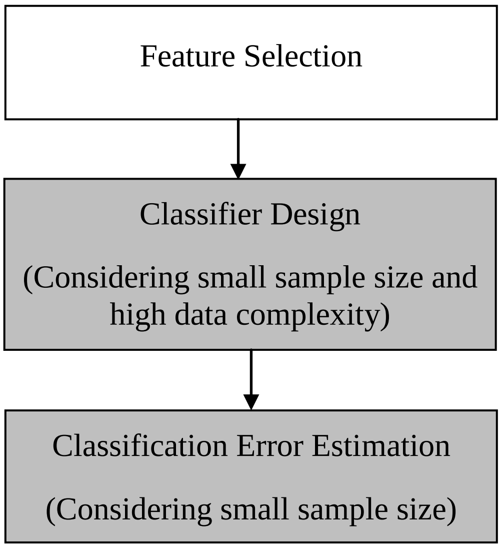
Feature Selection
Classifier Design
(Considering small sample size and high data complexity)
Classification Error Estimation
(Considering small sample size)

Figure 6.1: Stages involved in designing a pattern classifier. The stages focused on in this study are shown in grey.

The methods employed to achieve these aims and the simulation results were explained in the previous chapters. The current chapter discusses a novel application of PNN in breast cancer prediction. The PNN has been designed and developed during the period of this study based on-all the available markers and the criteria for achieving high accuracy. The breast cancer outcome in this thesis was chosen because of the significant role of the presence of tumour in the ALNs in medical decision making according to the medical literature [70]. And yet, previous research indicates the lack of agreement on a combination of predictive markers for accurate nodal status prediction that can replace the invasive procedures such as surgery and SLNB [7, 115].

Small sample size and complex nonlinear input-output relationship are common problems in predicting the ALN metastasis in breast cancer. Defining an adjuvant therapy is largely dependent on the prediction of nodal metastasis to the ALNs which is currently carried out through invasive procedures such as surgery or SLNB. Since many biomarkers, obtained in a non-invasive manner, have been identified to be related to the state of nodal involvement in breast cancer, it is advantageous to design a classifier which can classify patients using the non-invasively acquired data. For this purpose, the designed classifier should be able to capture the complex nonlinear input-output relation while avoiding overfitting to the available data. This necessitates designing the classifier with all available cases, and then to evaluate the generalization ability of the designed classifier using a resampling method.

In this study, the PNN has been employed as a platform to predict the state of nodal involvement in breast cancer patients based on tumour size and five biomolecular markers as well as their different combinations. An MLP neural network was also designed for this application as a benchmark method to be compared against the PNN results. The two ANNs' results were obtained by .632 bootstrap method for a reliable error estimation. For a clear investigation of results, the discussion in this chapter is divided into four parts. The reliability of error estimation is investigated in the first part. The two following parts contain the comparison of two ANNs and the analysis of results in terms of biomarkers' prediction potential respectively. The last part of discussion is dedicated to explain how the disadvantages in previous studies have been addressed in the current work.

6.2. Reliability of Network Evaluation

6.2.1. Bias

The bias of an error estimation method for a classifier is measured form the difference between its output and the true error. Computing the true error requires testing the network on the whole population. This is not possible for practical applications like DSSs since the available data are always limited. Nonetheless, it is possible to measure the bias of different error estimation methods by generating data assuming specific distributions for each class. In this way, the estimated and the true error of a classifier can be compared by testing the classifier for a limited and for a large sample size. This has been previously carried out by Sahiner et al. [28] to compare the bias of different error estimation methods. Their focus was on a particular application for medical diagnosis, with small sample sizes and the conclusion of the superiority of the resampling methods including .632 and .632+ bootstrap over other resampling methods.

It should be mentioned that .632 bootstrap can also suffer from optimistic bias in the case of strong overfitting (*0*% apparent error). Therefore, .632+ bootstrap was proposed to avoid this problem by adaptively adjusting the weights in .632 bootstrap [75]. However, .632+ bootstrap may result in overcorrecting the bias of the .632 bootstrap. Hence, .632 bootstrap is more reliably employed with methods with apparent errors of greater than *0*% [75]. In this study, apparent errors of higher than 5% were obtained where the best prediction results were achieved from the groups with apparent errors of higher than 8%.

6.2.2. Variance

For reliable comparison of the results obtained by the PNN and the MLP, it was important to use an accurate and reliable error estimation method. This is especially important in the application of ANNs in cancer prediction using patient biomarkers due to small sample sizes available in this field. 0.632 bootstrap was therefore employed to estimate the error of the two ANNs. This resampling technique was chosen for its low variance compared to the widely used cross validation method [116]. This was investigated in this study by comparing the variance of AUCs obtained by .632 bootstrap, 5-fold cross validation and holdout methods. It is evident from the results in Figure 5.1 to 5.4 that 5-fold cross validation and holdout methods have a significantly higher AUC variance than the .632 bootstrap AUC variance. The boxplots in these figures illustrate the high variability and the existence of outliers when computing the AUC using holdout method. This variability is reduced when using the 5-fold CV and is the lease with the use of .632 bootstrap. This is true of all combinations regardless of the employed network (MLP or PNN), input dimension (Figures 5.1 and 5.3) or the input markers (Figures 5.2 and 5.4).

The .632 bootstrap low variance results from the fact that the randomness in CV training and test sets is eliminated by using 50 bootstrap replications of the training data. In addition, misclassification error rate computed using .632 bootstrap is fairly unbiased and hence more reliable compared to apparent error estimator which suffers from low bias [28, 75, 117]. The low bias in apparent error happens as the same data employed for training the network is then used to test it.

6.3. Artificial neural network performance analysis

The MLP and the PNN are both trained in a supervised manner (i.e. the corresponding output of each input pattern is available to the network during the training course). The fundamental difference between the MLP and the PNN lies in their different data analysis approaches applied to perform the classification task. This can be represented by the dichotomous statistical pattern classification approaches denoted by Jain et al [118] as “Density-Based” and “Geometric” approaches. These methods discriminate based on whether the decision boundary is derived indirectly from the PDF estimation of training data or directly from the training data. The PNN data analysis approach to perform the classification task can be represented by the “Density-Based” statistical pattern classification approach. This method discriminates based on deriving the decision boundary indirectly from the PDF estimation of training data. Hence, in the PNN data classification approach, approximating the underlying distribution of data by estimating the PDF is targeted. Alternatively, data classification in the MLP is geometric which is carried out by handling the numerical data directly. The MLP performs pattern classification by constructing the decision boundaries based on the training data and optimising a cost function.

MLP has been widely used in cancer studies in spite of some common drawbacks such as network complexity, optimisation of numerous network variables and random weight initialisation. The PNN maintains unique advantages which makes it an alternative ANN structure to the MLP for some classification problems. These advantages include:

- its rapid training capability which makes it many times faster than the SCG algorithm and guaranteed convergence to the optimal Bayes classifier by choosing the most probable classification among all possible classes;
- it is easily modified by adding new training data and hence is compatible with online applications, unlike the SCG algorithm which needs to be retrained for any modification in the training data;
- the PNN outputs are interpretable in the sense that the inputs can be characterized by their effectiveness in making the output decision [119].

On the other hand, the PNN requires a large amount of memory to run, as all the training data needs to be stored during the PNN training process [24]. This is not an issue in small sample size classification problems as the number of available data for training is limited and hence, the amount of required memory to store the data does not pose a problem.

In addition, the prior knowledge is employed in the PNN in combination with the data to classify a new input. This prior knowledge is in the form of the relative frequency of the output categories in the data.

PNN also addresses two main practical difficulties in the Bayesian methods. The first difficulty in Bayesian methods is the requirement of having knowledge about the data distributions. In classification problems, these distributions are not known and they must be estimated from the available data. The conventional statistical classification methods estimate these probabilities in a parametric manner, i.e. by making assumptions about the form of the underlying distributions. Parametric density estimation methods assume a unimodal density distribution with one local maximum. However, most practical classification problems involve data with multimodal densities. In order to model densities with more than one local maximum, PDFs can be estimated in a nonparametric manner. Nonparametric density estimation methods make no assumption about the underlying distribution of the data and therefore, they can be exercised for arbitrary distributions [44].

Another advantage of the PNN over the MLP is the use of prior probability. The estimation of prior in classification problems can be simply obtained from the relevant frequency of each category.

Generally, the PNN performed more accurately than the MLP. The mean accuracy of 19 chosen combinations obtained by the PNN was more than the mean obtained by MLP by 4%. PNN attained the best accuracy of 69%, the best sensitivity of 67% and specificity of 66%.

Better results obtained by PNN compared to MLP can be explained by the fundamental characteristic of the learning procedures in the PNN and the MLP. Both network structures use supervised learning wherein the aim is to obtain an optimum set of weights in the network that best describes the relationship between a set of input-output patterns and can be generalized for predicting new patterns. Learning in the PNN was carried out by employing prior knowledge about the data population and defining weights using training data. This prior knowledge was implemented in the PNN in the form of the ratio of the frequency of training samples from each class in addition to modelling the probability densities of the training patterns using

Gaussian kernel. Classification in PNN was then performed using the Bayesian approach to describe the relationship between input-output patterns by computing the posterior probability. In the MLP, training was carried out by minimising a cost function derived from the training data based on maximum likelihood approach. Therefore, learning in the MLP included an iterative procedure, as described in section 4.2., which uses only the information provided by the training data to define an optimum set of weights. Accordingly, PNN has an advantage over the MLP in that it uses both prior knowledge and training data to achieve classification.

Comparing the output apparent error of different marker combinations, it was observed that although the combination including all markers obtained the lowest apparent error of 5% in the PNN, its test accuracy was not as high as some of the other combinations. The fact that the inclusion of all the variables offered the best design accuracy indicated that each marker makes a contribution to the classification process and yet some specific patterns of relationship emerge between the input and outputs that lead to a poor prediction, which probably reflects the relative significance of the variables engendered by the interaction between them.

6.4. Predictive Value of Markers

Establishing the efficacy of the markers through grouping can provide good indications of the relative contribution of each marker to outcome prediction and the relationships it may have with other potential markers.

Several important and novel features' combinations have emerged from the present investigation. Although tumour markers have been used for many years, an ANN seems capable of not only judging the weights or significance of individual makers to disease progression, but also, as shown in this study, it provides a decision making capability that takes into account the interactions between the signalling by different markers that indeed governs the route of tumour progression.

With the PNN neural network, the best prediction accuracy of 69% was obtained by a combination of ER, p53, Ki-67 and age with and without PR. Steroid receptor status and the degree of p53 expression or other proliferation markers have often offered fertile grounds for debate as to whether they could serve as independent markers of cancer progression and prognosis. Historically, steroid hormone receptor status has received much attention from the clinical and research faculty in this regard. The present work sheds light on the relevance of ER/PR in assessing cancer progression. The interaction between signalling by the steroid hormones through their respective receptors is evident from the finding in this study that omission of ER or PR individually does not make a sizeable effect in the predictive ability. Since ER in an active functional state can induce the expression of PR, it follows that the presence of either marker would suffice to provide the biological factor required for asserting predictive ability. Omission of both ER and PR markedly reduces the prediction scores in PNN. Overall, the conservative conclusion is that ER/PR might influence prediction as independent factors.

Individually, the markers have low predictive ability in the range of 51-61% in PNN. It is the success achieved by combining different markers which deserves special emphasis here. Tumour size appears from this analysis to be an important contributor to the predictive ability, for its omission resulted in a marked reduction in predictive accuracy. This is possibly related to the higher invasive ability of larger tumours which attributes to the greater proliferative pressure within the tumour.

The cell cycle regulator p53 is mutated or lost in a majority of human cancers. Hence the debate has addressed issues relating to the p53 expression as an aid in assessing tumour behaviour. The results of the study show that the omission of p53 and/or Ki-67 leads to approximately 5% loss in predictive accuracy. This could indicate the possibility that p53 might not be an independent marker of progression.

Finally, one important line of reasoning that this investigation seems to generate and highlight is whether ER/PR signalling interacts with the p53 signalling pathway. As noted earlier, ER and PR might be able to compensate for omission of p53, but the question that needs to be answered is whether ER and PR make an independent contribution to the process of tumour progression. An early approach to this question was to test if resistance to tamoxifen was related to p53 over expression, although refractoriness of breast cancer cell lines to tamoxifen can be due to other factors such as the expression of HER2. In breast cancer, p53 did not appear to be significantly associated with response to tamoxifen, although tumours over expressing p53 protein were more aggressive and the patients showed poorer survival [120]. Loss of ER brought about by methylation in 26% and abnormalities of p53 pathway in 53% of cases was found in an investigation to discover whether in endometrial cancers ER and p53 pathways were inter-related [121]. However, there was no discernible link between the loss of ER and abnormalities in p53 signalling. On the other hand, others have suggested some correlation between p53 and ER signalling. Ovarian tumours have been reported to show abnormalities in mdm2 and p14ARF, both related to p53 signalling, and abnormalities. ER has been implicated in the regulation of this p14ARF-MDM2-p53 pathway [122]. Kang [123] has reported that mutations of p53 results in increased expression of pS2, a downstream target of ER. This effect could be abolished by inhibiting a p53-directed signalling system involving the extracellular signal-regulated kinases.

The best prediction obtained from the combination of all markers excluding tumour size maintains a low error test while upholding reasonable apparent error. Moreover, this subset showed a sound value for sensitivity which was very low in other combinations. Therefore, this subset was used for comparison with other subsets. Omission of age from this best performing set markedly increased the apparent error, albeit with reduction also in the test accuracy. In general, when age is excluded, the accuracy results are lower than when it is included. In relation to the other markers, age produced strong prediction results.

6.5. How disadvantages in the previous models are addressed

ANNs have been employed in the field of medical diagnosis for nearly 20 years . They are suitable for both classification and regression, are tolerant to noisy inputs and their structure and connection weights can be configured such that they will be able to represent Boolean functions (AND, OR, NOT). However, the widespread use of them is sometimes criticized due to some common drawbacks such as the limitation of understanding the algorithm structure constructed by the network connections, large number of parameters which may result in overfitting and difficulty in defining an optimal network structure (optimal number of hidden layers and nodes) [50].

In addition, the network must be able to predict the outcome for new data which might be slightly different from the training data. Therefore, constructing the network such that it can only predict the training data perfectly is not desirable. In order to estimate the generalization ability of the network, it is important to estimate the error of the designed ANN for new data. However, reliable error estimation under the constraint of small sample size is another issue that has been neglected in many ANN-based studies for cancer prediction. Below is a discussion of how these drawbacks have been addressed in this study.

6.5.1. Overfitting

Overfitting is one of the main issues in training the MLPs. In practical applications and especially in the field of medical diagnosis, the dataset includes limited number of samples. Additionally, this should be further partitioned for the training and test purposes to provide independent evaluation of network ability for the prediction of new samples. MLPs normally perform well for predicting the training data but they provide poorer predictions for the test data. The poor generalisation ability of the ANNs results from overfitting the network parameters to the training data.

As mentioned in chapter 2, CV can be employed to detect when the network starts to overfit during the training process and hence, avoiding the overfitting by discontinuing the training process from that point. However, CV is not an optimum method as a part of the data is kept for the validation purpose and hence, the network is trained only with a part of data. This is especially problematic with small datasets as the patterns in the data must be unveiled only with a part of the small data.

In this study, the overfitting in the MLPs was obviated by employing a PNN structure which maintains small number of free parameters to be estimated. Among different supervised ANN structures, PNN has the least number of free parameters with only one parameter to estimate. This characteristic allows the PNNs to be employed in conjunction with small data sizes while the MLP was limited in this manner. The large number of free parameters in MLPs complicates choosing an optimised network structure for the problem at hand. Main free parameters of MLPs include number and size of hidden layers, hidden layer transfer function, cost function employed for training the network and initial weights and biases. Unfortunately, there is no agreement on a definite method for choosing an optimum amount for these parameters. In most studies, these parameters are decided by experiment [118]. Nevertheless, the PNN's only free parameter, the spread, was defined in this study by Silverman's rule thumb that has been proven to provide a good approximate for the spread of the data while evading over fitting to the training data [124].

6.5.2. Error estimation

Most ANN applied to diagnostic classification problems suffer from biased estimation of the classification error [20]. The random division of data (holdout method) has been used in most studies as an error estimation method for MLP. However, holdout method results in a high variance in the output results and can become misleading by reporting only the best achieved results in multiple runs of network with different random selections of training and test sets. On the other hand, unbiased estimation of error via cross validation results in the high variability of estimated errors by differently partitioning the data. In this study, a reliable estimation of network classification error was achieved by employing .632 bootstrap method. In addition, this method maintains a low bias which makes it an apt choice to be employed in conjunction with small medical data sizes. Moreover, the limitation of small sample sizes can be obviated using .632 bootstrap resampling method to generate new training and test sets with the same distribution as the original data.

6.6. Summary

The main three sections of results presented in the previous chapter were discussed in this chapter in detail. The first section covered the comparison between different error estimation methods and the reasoning for the superiority of .632 bootstrap over CV and holdout methods. The comparison results of the PNN and the MLP were elaborated in the second part of this chapter. While both methods maintain some unique advantages, the PNN was more suitable for the application of ANL metastasis prediction and performed more accurately than the MLP. The predictive potential of the markers together with some medical literature were also discussed. The best prediction was obtained from the combination of all markers excluding tumour size which demonstrates that including more markers does not necessarily lead to higher accuracies. Some of the gaps in the application of ANNs in cancer studies and how they were addressed in this research were covered in the last section of this chapter. The next chapter entails the main conclusions of the study and how it contributes to a wider context of research.

Chapter 7

7. Conclusions and future work

7.1. Research contributions

Much research has been devoted to the study of ANNs' application in cancer prediction. In addition, there is background literature available on the reliability of resampling error estimation methods used in conjunction with ANNs [28]. However, this is the first study of its kind that investigates the application of PNN in the prediction of ALN metastasis using a reliable error estimation method to obviate the issue of small sample size. The developed approach is capable of synthesizing and integrating specific biomarkers of the breast tumour in order to carry out complex evaluations and present the results to the clinicians in a timely manner. Moreover, this study offers explanation, comparison and performance assessment of two feedforward neural networks with application of predicting the state of nodal involvement in breast carcinoma.

Three conclusions can be drawn from the present series of experiments, one relating to the reliability of ANNs in conjunction with small datasets. The second is regarding the networks' architectures and the analysis of PNN and MLP potential for the present application. The third is regarding the markers' significance in cancer prognosis.

Generally, ANNs offer the possibility to investigate the complex relationships between individual markers affecting tumour progression and prognosis. However, a judicious choice of an error estimation method for the evaluation of ANN outcome leads to significant improvement in the reliability of outcome analysis. From the results and comparisons in this study, it is clear that the .632 bootstrap method gives a reliable validation of network accuracy compared to holdout and 5-fold CV. Besides, the .632 bootstrap approach makes it viable to exploit all the available data for training purpose that is especially advantages in medical studies with the constraint of small sample size. High error estimation reliability offered by .632 bootstrap allows ANN systems to maintain a good generalization and provide reliable analysis in limited datasets.

Among the two ANN structures considered in the study, the results show that for the present application, the PNN performs better than the MLP by providing better prediction accuracy, especially when a large number of markers are present. The main disadvantage of the PNN over MLP is that it requires a large number of hidden units as the size of hidden layer in the PNN depends on the number of input patterns. This drawback is negligible in most medical applications as the datasets in this domain usually contain only a limited number of patterns due to the intricacy in measuring the biomarkers. On the other hand, the training time required by the PNN is considerably shorter than MLP. Although the SCG algorithm employed for training the MLP is a relatively fast method compared with the BP algorithm, it still needs several iterations to find the minimum of the error surface and provide an optimum output prediction. The PNN method only needs a single forward iteration for the training process which makes it a considerably faster method than the MLP. An overview of the two ANN structures has led to the conclusion that in general terms, the PNN network provides better prediction outcome while it seems to uphold the major findings with the MLP network in terms of the importance of tumour size and p53 as factors that influence tumour progression.

With regard to the markers' prognostic significance, the full set of markers is required to provide the highest classification accuracy in the training process. However, the subset including ER, p53, Ki-67 and age results in a better test accuracy for both the MLP and the PNN. Although individual markers achieve low predictive scores, certain combinations do indicate marked benefit in terms of prediction of nodal involvement. Combination of tumour size appears from this analysis to be an important contributor to the predictive ability possibly reflecting the higher invasive ability of larger tumours. The omission of Ki-67 expression results in a slight loss in predictive accuracy suggesting Ki-67 might not be an independent marker of progression.

The present work has clearly demonstrated that the steroid signalling might be compensating for the omission of p53/Ki-67, for the omission of both ER and PR markedly reduces predictive accuracy. This confirms the importance of ER and PR in assessing progression and also provides evidence for the interaction between signalling by these steroids. The present work suggests that ER/PR might compensate for the omission of p53, and further there are clear indications here that ER and p53 do not greatly affect each other's influence on predictive outcome and therefore on disease progression.

It is recommend that the classification accuracy of this panel of markers possesses significant potential for predicting nodal spread of breast cancer in individual patients. The accuracy which seems to flow from the biological features that define disease progression should greatly reduce the need for adopting invasive procedures, and in this way reduce the post-operative stress and associated morbidity experience by patients. Furthermore, it is needless to reiterate the potential reduction of the time lag between investigation and decision making in patient management.

7.2. Future work

This study meets the criteria for a study design in medical diagnosis including a good structure and reliable validation. Even so, there are certain aspects of the study that can be improved given more time and resources. These mainly include employing the PNN results for further interpretation and utilising a larger dataset for the design and validation of the study.

- The present work has clearly demonstrated the superiority of the PNN over conventional statistical and machine learning methods and also the commonly employed MLP. This superiority comes from the special properties of the PNN including fast training, easily modifiable structure and meeting Bayesian optimality in classification. In this study, these unique characteristics are proved to significantly improve the accuracy and reliability of breast cancer prediction. As mentioned before, the improvement in the reliability is achieved by employing the bootstrap evaluation which is possible to apply in practice due to the fast training of the PNN. One important characteristic of the PNN which has been mentioned in this study is its interpretability in terms of the output probability. Using the output probabilities for further interpretation of the results is a future aim of the study as it may shed more light on the potential of each biomarker in prediction.

- Regarding the breast cancer data, as the availability of data is limited in this domain, collection of more data for further substantiating the results would be sought in future. Upon validating the results with a larger sample of patients, the proposed techniques would be applicable in practice for breast cancer diagnosis. In addition to the size of the data, the performance of ANN largely depends on the measurement accuracy of the employed markers for classification. Those biomarkers which need a human observer for quantification are prone to provide poor data quality given an inexpert observer. Therefore, obtaining all biomarkers in an automatic fashion is sought in the future studies.

References

[1] "Breast Cancer in England, 2009," in *Office for National Statistics*, 2011, URL: http://www.ons.gov.uk/ons/rel/cancer-unit/breast-cancer-in-england/2009/breast-cancer.html.

[2] R. Peto, J. Boreham, M. Clarke, C. Davies, and V. Beral, "UK and USA breast cancer deaths down 25% in year 2000 at ages 20-69 years," *The Lancet*, vol. 355, pp. 1822-1822, 2000.

[3] S. Samphao, J. M. Eremin, M. El-Sheemy, and O. Eremin, "Management of the Axilla in Women With Breast Cancer: Current Clinical Practice and a New Selective Targeted Approach," *Annals of Surgical Oncology*, vol. 15, pp. 1282-1296, 2008.

[4] B. Fisher, M. Bauer, D. L. Wickerham, C. K. Redmond, and E. R. Fisher, "Relation of number of positive axillary nodes to the prognosis of patients with primary breast cancer. An NSABP update," *Cancer*, vol. 52, pp. 1551-1557, 1983.

[5] Y. Takatsuka, "Prediction of lymph node metastases in breast cancer by clinicopathological and biological features of the primary tumor," *Breast Cancer*, vol. 6, pp. 155-158, 1999.

[6] M. Keshtgar, N. Aresti, and F. Macneil, "Establishing axillary Sentinel Lymph Node Biopsy (SLNB) for early breast cancer in the United Kingdom: a survey of the national training program," *Eur J Surg Oncol*, vol. 36, pp. 393-8, 2010.

[7] N. R. Patani, M. V. Dwek, and M. Douek, "Predictors of axillary lymph node metastasis in breast cancer: A systematic review," *European Journal of Surgical Oncology,* vol. 33, pp. 409-419, 2007.

[8] B. Fisher, R. G. Ravdin, R. K. Ausman, N. H. Slack, G. E. Moore, and R. J. Noer, "Surgical adjuvant chemotherapy in cancer of the breast: results of a decade of cooperative investigation," *Ann. Surg.,* vol. 168, pp. 337-56, Sep 1968.

[9] S. M. Veronese, M. Gambacorta, O. Gottardi, F. Scanzi, M. Ferrari, and P. Lampertico, "Proliferation index as a prognostic marker in breast cancer," *Cancer,* vol. 71, pp. 3926-3931, 1993.

[10] F. J. Esteva and G. N. Hortobagyi, "Prognostic molecular markers in early breast cancer," *Breast Cancer Res,* vol. 6, pp. 109-18, 2004.

[11] M. Andronas, S. S. Dlay, and G. V. Sherbet, "Oestrogen and progesterone receptor expression influences DNA ploidy and the proliferation potential of breast cancer cells," *Anticancer Research,* vol. 23, pp. 3029-3039, 2003.

[12] H. Seker, M. O. Odetayo, D. Petrovic, R. N. G. Naguib, C. Bartoli, L. Alasio, M. S. Lakshmi, and G. V. Sherbet, "Soft feature evaluation indices for the identification of significant image cytometric factors in assessment of nodal involvement in breast cancer patients," in *Fuzzy Systems. FUZZ-IEEE'02. Proceedings of the IEEE International Conference on,* 2002, pp. 1592-1595.

[13] H. Seker, M. Odetayo, D. Petrovic, R. N. G. Naguib, C. Bartoli, L. Alasio, M. S. Lakshmi, and G. V. Sherbet, "Prognostic comparison of statistical, neural and fuzzy methods of analysis of breast cancer image cytometric data," in *Engineering in Medicine and Biology Society. Proceedings of the*

23rd Annual International Conference of the IEEE, 2001, pp. 3811-3814 vol.4.

[14] H. B. Burke, D. B. Rosen, and P. H. Goodman, "Comparing artificial neural networks to other statistical methods for medical outcome prediction," in *Neural Networks. IEEE World Congress on Computational Intelligence., IEEE International Conference on*, 1994, pp. 2213-2216.

[15] H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. Harrell, J. J. R. Marks, D. P. Winchester, and D. G. Bostwick, "Artificial neural networks improve the accuracy of cancer survival prediction," *Cancer,* vol. 79, pp. 857-862, 1997.

[16] G. D. Tourassi, C. E. Floyd, Jr., and J. Y. Lo, "A constraint satisfaction neural network for medical diagnosis," in *Neural Networks, 1999. IJCNN '99. International Joint Conference on*, 1999, pp. 3632-3635 vol.5.

[17] R. N. G. Naguib, H. A. M. Sakim, M. S. Lakshmi, V. Wadehra, T. W. J. Lennard, J. Bhatavdekar, and G. V. Sherbet, "DNA ploidy and cell cycle distribution of breast cancer aspirate cells measured by image cytometry and analyzed by artificial neural networks for their prognostic significance," *Information Technology in Biomedicine, IEEE Transactions on,* vol. 3, pp. 61-69, 1999.

[18] V. S. Bourdes, S. Bonnevay, P. J. G. Lisboa, M. S. H. Aung, S. Chabaud, T. Bachelot, D. Perol, and S. Negrier, "Breast Cancer Predictions by Neural Networks Analysis: a Comparison with Logistic Regression," in *Engineering in Medicine and Biology Society. EMBS 2007. 29th Annual International Conference of the IEEE*, 2007, pp. 5424-5427.

[19] M. L. Astion and P. Wilding, "Application of neural networks to the interpretation of laboratory data in cancer diagnosis," *Clin Chem,* vol. 38, pp. 34-38, 1992.

[20] G. Schwarzer, W. Vach, and M. Schumacher, "On the misuses of artificial neural networks for prognostic and diagnostic classification in oncology," *Statistics in Medicine,* vol. 19, pp. 541-561, 2000.

[21] F. E. Ahmed, "Artificial neural networks for diagnosis and survival prediction in colon cancer," *Mol Cancer*, vol. 4, p. 29, 2005.

[22] S. R. Grey, S. S. Dlay, G. V. Sherbet, B. E. Leone, and F. Cajone, "Artificial neural network assessment of the prognostic value of tumour promoter gene S100A4 and metastasis suppressor gene nm23 in conjunction with tumour grade, size and steroid receptor status of breast cancer," *AACR Meeting Abstracts*, pp. 1109-b, 2004.

[23] D. F. Specht, "Probabilistic neural networks for classification, mapping, or associative memory," in *Neural Networks, 1988., IEEE International Conference on*, 1988, pp. 525-532

[24] D. F. Specht, "Probabilistic neural networks," *Neural Networks,* vol. 3, pp. 109-118, 1990.

[25] C. J. Huang and W. C. Liao, "A comparative study of feature selection methods for probabilistic neural networks in cancer classification," in *Tools with Artificial Intelligence. Proceedings. 15th IEEE International Conference on*, 2003, pp. 451-458.

[26] A. N. Karahaliou, I. S. Boniatis, S. G. Skiadopoulos, F. N. Sakellaropoulos, N. S. Arikidis, E. A. Likaki, G. S. Panayiotakis, and L. I. Costaridou, "Breast Cancer Diagnosis: Analyzing Texture of Tissue Surrounding

Microcalcifications," *Information Technology in Biomedicine, IEEE Transactions on,* vol. 12, pp. 731-738, 2008.

[27] Y. Shan, R. Zhao, G. Xu, H. M. Liebich, and Y. Zhang, "Application of probabilistic neural network in the clinical diagnosis of cancers based on clinical chemistry data," *Analytica Chimica Acta,* vol. 471, pp. 77-86, 2002.

[28] B. Sahiner, H. P. Chan, and L. Hadjiiski, "Classifier performance estimation under the constraint of a finite sample size: Resampling schemes applied to neural network classifiers," *Neural Networks,* vol. 21, pp. 476-483, 2008.

[29] P. Mitra, S. Mitra, and S. K. Pal, "Staging of cervical cancer with soft computing," *Biomedical Engineering, IEEE Transactions on,* vol. 47, pp. 934-940, 2000.

[30] R. Setiono, B. Baesens, and C. Mues, "Recursive Neural Network Rule Extraction for Data With Mixed Attributes," *Neural Networks, IEEE Transactions on,* vol. 19, pp. 299-307, 2008.

[31] R. S. Ledley and L. B. Lusted, "Reasoning Foundations of Medical Diagnosis," *Science,* vol. 130, pp. 9-21, 1959.

[32] R. Dybowski and V. Gant, "Artificial neural networks in pathology and medical laboratories," *The Lancet,* vol. 346, pp. 1203-1207, 1995.

[33] J. F. Kenney and E. S. Keeping, "Linear Regression and Correlation," in *Mathematics of Statistics, Pt. 1,* 3rd ed Princeton, NJ: Van Nostrand, 1962, pp. 252-285.

[34] A. F. Cabrera, "Logistic regression analysis in higher education: An applied perspective," in *Higher Education: Handbook of Theory and Research*. vol. 10, 1994, pp. 225–56.

[35] P. J. G. Lisboa, H. Wong, P. Harris, and R. Swindell, "A Bayesian neural network approach for modelling censored data with an application to prognosis after surgery for breast cancer," *Artificial Intelligence in Medicine,* vol. 28, pp. 1-25, 2003.

[36] P. J. G. Lisboa, "A review of evidence of health benefit from artificial neural networks in medical intervention," *Neural Networks,* vol. 15, pp. 11-39, 2002.

[37] R. Bittern, A. Cuschieri, S. G. Dolgobrodov, R. Marshall, and P. Moore, "Artificial Neural Networks in Cancer Management."

[38] E. Biganzoli, P. Boracchi, L. Mariani, and E. Marubini, "Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach," *Stat. in Med.,* vol. 17, pp. 1169-86, 1998.

[39] D. W. Hosmer and S. Lemeshow, *Applied logistic regression*, 2nd ed.: John Wiley and Sons, 2000.

[40] K. Van Zee, D.-M. Manasseh, J. Bevilacqua, S. Boolbol, J. Fey, L. Tan, P. Borgen, H. Cody, and M. Kattan, "A Nomogram for Predicting the Likelihood of Additional Nodal Metastases in Breast Cancer Patients With a Positive Sentinel Node Biopsy," *Annals of Surgical Oncology,* vol. 10, pp. 1140-1151, 2003.

[41] S. J. Press and S. Wilson, "Choosing Between Logistic Regression and Discriminant Analysis," *Journal of the American Statistical Association,* vol. 73, pp. 699-705, 1978.

[42] K. Fukunaga and J. M. Mantock, "Nonparametric Discriminant Analysis," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. PAMI-5, pp. 671-678, 1983.

[43] S. S. Cross, R. F. Harrison, and R. L. Kennedy, "Introduction to neural networks," *The Lancet,* vol. 346, pp. 1075-1079, 1995.

[44] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification*: Wiley, John & Sons, 2001.

[45] F. Murtagh, "Multilayer perceptrons for classification and regression," *Neurocomputing,* vol. 2, pp. 183-197, 1991.

[46] G. P. Zhang, "Neural networks for classification: a survey," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* vol. 30, pp. 451-462, 2000.

[47] E. Gamito and E. Crawford, "Artificial neural networks for predictive modeling in prostate cancer," *Current Oncology Reports,* vol. 6, pp. 216-221, 2004.

[48] J. A. Gómez-Ruiz, J. M. Jerez-Aragonés, J. Muñoz-Pérez, and E. Alba-Conejo, "A Neural Network Based Model for Prognosis of Early Breast Cancer," *Applied Intelligence,* vol. 20, pp. 231-238, 2004.

[49] T. M. Mitchell, *Machine Learning*: WCB/McGraw-Hill, 1997.

[50] J. A. Cruz and D. S. Wishart, "Applications of machine learning in cancer prediction and prognosis," *Cancer Inform,* vol. 2, pp. 59-77, 2006.

[51] Y. Su, J. Shen, H. Qian, H. Ma, J. Ji, L. Ma, W. Zhang, L. Meng, Z. Li, J. Wu, G. Jin, J. Zhang, and C. Shou, "Diagnosis of gastric cancer using decision tree classification of mass spectral data," *Cancer Sci,* vol. 98, pp. 37-43, Jan 2007.

[52] A. Vlahou, J. O. Schorge, B. W. Gregory, and R. L. Coleman, "Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data," *J Biomed Biotechnol,* vol. 2003, pp. 308-314, 2003.

[53] E. Papageorgiou, C. Stylios, and P. Groumpos, "A Combined Fuzzy Cognitive Map and Decision Trees Model for Medical Decision Making," in *Engineering in Medicine and Biology Society, 2006. EMBS '06. 28th Annual International Conference of the IEEE,* 2006, pp. 6117-6120.

[54] M. U. Khan, J. P. Choi, H. Shin, and M. Kim, "Predicting breast cancer survivability using fuzzy decision trees for personalized healthcare," *Conf Proc IEEE Eng Med Biol Soc,* vol. 2008, pp. 5148-51, 2008.

[55] J. R. Quinlan, "Induction of Decision Trees," *Machine Learning* vol. 1, pp. 81-106, 1986.

[56] J. Bray, J. Sludden, M. J. Griffin, M. Cole, M. Verrill, D. Jamieson, and A. V. Boddy, "Influence of pharmacogenetics on response and toxicity in breast cancer patients treated with doxorubicin and cyclophosphamide," *Br J Cancer,* vol. 102, pp. 1003-1009.

[57] W. S. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *Bull Math Biol,* vol. 5, pp. 115–133 1943.

[58] R. Linder, I. R. Konig, C. Weimar, H. C. Diener, S. J. Poppl, and A. Ziegler, "Two models for outcome prediction - A comparison of logistic regression and neural networks," *Methods of Information in Medicine,* vol. 45, pp. 536-540, 2006.

[59] D. J. Sargent, "Comparison of artificial neural networks with other statistical approaches," *Cancer,* vol. 91, pp. 1636-1642, 2001.

[60] D. Michie, D. J. Spiegelhalter, and C. C. Taylor, *Machine Learning, Neural and Statistical Classification* Ellis Horwood 1994.

[61] B. Kosko, "Unsupervised learning in noise," *Neural Networks, IEEE Transactions on,* vol. 1, pp. 44-57, 1990.

[62] S. Haykin, *Neural Networks and Learning Machines*, Third ed.: Prentice Hall, 2009.

[63] D. Lowe and A. Webb, "Exploiting prior knowledge in network optimization: an illustration from medical prognosis," *Network: Computation in Neural Systems,* vol. 1, pp. 299-323, 1999.

[64] B. D. Ripley, *Pattern Recognition and Neural Networks*. Cambridge: Cambridge University Press, 1996.

[65] C. M. Bishop, *Neural networks for pattern recognition*: Oxford. University Press, 1995.

[66] F. Rosenblatt, "The perceptron: a probabilistic method for information storage in the brain," *Psych Rev,* vol. 65, pp. 386-407, 1959.

[67] R. N. G. Naguib and G. V. Sherbet, *Artificial Neural Networks in Cancer Diagnosis, Prognosis, and Patient management*: CRC Press

2001.

[68] W. G. Baxt, "Application of artificial neural networks to clinical medicine," *The Lancet,* vol. 346, pp. 1135-1138, 1995.

[69] R. N. G. Naguib and G. V. Sherbet, "Artificial Neural Networks in Cancer Research," *Pathobiology,* vol. 65, pp. 129-139 1997.

[70] T. Mattfeldt, H. A. Kestler, and H. P. Sinn, "Prediction of the axillary lymph node status in mammary cancer on the basis of clinicopathological data and flow cytometry," *Med Biol Eng Comput,* vol. 42, pp. 733-9, 2004.

[71] S. R. Grey, S. S. Dlay, B. E. Leone, F. Cajone, and G. V. Sherbet, "Prediction of nodal spread of breast cancer by using artificial neural network-based analyses of S100A4, nm23 and steroid receptor expression," *Clinical and Experimental Metastasis,* vol. 20, pp. 507-514, 2003.

[72] R. N. G. Naguib, A. E. Adams, C. H. W. Horne, B. Angus, G. V. Sherbet, and T. W. J. Lennard, "The detection of nodal metastasis in breast cancer using neural network techniques," *Physiological Measurement,* vol. 17, pp. 297-303, 1996.

[73] S. A. Mojarad, S. S. Dlay, W. L. Woo, and G. V. Sherbet, "Breast cancer prediction and cross validation using multilayer perceptron neural networks," in *7th IEEE, IET International Symposium on Communication Systems Networks and Digital Signal Processing (CSNDSP),* 2010, pp. 760-764.

[74] J. C. Wyatt and D. G. Altman, "Commentary: Prognostic models: clinically useful or quickly forgotten?," *BMJ,* vol. 311, pp. 1539-1541, 1995.

[75] B. Efron and R. Tibshirani, "Improvements on Cross-Validation: The .632+ Bootstrap Method," *Journal of the American Statistical Association,* vol. 92, pp. 548-560, 1997.

[76] B. Efron and R. Tibshirani, *Introduction to the Bootstrap* Chapman & Hall/CRC Monographs on Statistics & Applied Probability, 1993.

[77] B. Gerber, K. Heintze, J. Stubert, M. Dieterich, S. Hartmann, A. Stachs, and T. Reimer, "Axillary lymph node dissection in early-stage invasive breast cancer: is it still standard today?," *Breast Cancer Research and Treatment,* pp. 1-12, 2011.

[78] S. L. Wong, C. Chao, M. J. Edwards, D. J. Carlson, A. Laidley, R. D. Noyes, T. McGlothin, P. B. Ley, T. Tuttle, M. Schadt, R. Pennington, M. Legenza, J. Morgan, and K. M. McMasters, "Frequency of sentinel lymph node metastases in patients with favorable breast cancer histologic subtypes," *The American Journal of Surgery,* vol. 184, pp. 492-498, 2002.

[79] W. L. Donegan, "Tumor-related prognostic factors for breast cancer," *CA: A Cancer Journal for Clinicians*, vol. 47, pp. 28-51, 1997.

[80] C. L. Carter, C. Allen, and D. E. Henson, "Relation of tumor size, lymph node status, and survival in 24,740 breast cancer cases," *Cancer*, vol. 63, pp. 181-7, 1989.

[81] P. Valagussa, G. Bonadonna, and U. Veronesi, "Patterns of relapse and survival following radical mastectomy. Analysis of 716 consecutive patients," *Cancer*, vol. 41, pp. 1170-8, 1978.

[82] C. K. Osborne, "Steroid hormone receptors in breast cancer management," *Breast Cancer Research and Treatment*, vol. 51, pp. 227-238, 1998.

[83] R. Lapidus, S. Nass, and N. Davidson, "The loss of estrogen and progesterone receptor gene expression in human breast cancer," *J Mammary Gland Biol Neoplasia*, vol. 3, pp. 85-94, 1998.

[84] K. B. Horowitz and W. L. McGuire, "Predicting response to endocrine therapy in human breast cancer: a hypothesis," *Science*, vol. 189, pp. 726-727, 1975.

[85] C. Ballaré, M. Uhrig, T. Bechtold, E. Sancho, M. Di Domenico, A. Migliaccio, F. Auricchio, and M. Beato, "Two domains of the progesterone receptor interact with the estrogen receptor and are required for progesterone activation of the c-Src/Erk pathway in mammalian cells," *Mol Cell Biol.* , vol. 23, pp. 1994-2008, 2003.

[86] B. Chen, H. Pan, L. Zhu, Y. Deng, and J. W. Pollard, "Progesterone Inhibits the Estrogen-Induced Phosphoinositide 3-Kinase->AKT->GSK-3{beta}->Cyclin D1->pRB Pathway to Block Uterine Epithelial Cell Proliferation," *Mol Endocrinol*, vol. 19, pp. 1978-1990, 2005.

[87] P. D. P. Pharoah, N. E. Day, and C. Caldas, "Somatic mutations in the p53 gene and prognosis in breast cancer: a meta-analysis," *Br J Cancer,* vol. 80, pp. 1968-1973, 1999.

[88] P. de Cremoux, A. Vincent Salomon, S. Liva, R. m. Dendale, B. Bouchind'homme, E. Martin, X. Sastre-Garau, H. Magdelenat, A. Fourquet, and T. Soussi, "p53 Mutation as a Genetic Trait of Typical Medullary Breast Carcinoma," *Journal of the National Cancer Institute,* vol. 91, pp. 641-643, April 7, 1999 1999.

[89] K. H. Vousden and D. P. Lane, "p53 in health and disease," *Nat Rev Mol Cell Biol,* vol. 8, pp. 275-283, 2007.

[90] J. Tommiska, H. Eerola, M. Heinonen, L. Salonen, M. Kaare, J. Tallila, A. Ristimaki, K. von Smitten, K. Aittomaki, P. Heikkila, C. Blomqvist, and H. Nevanlinna, "Breast Cancer Patients with p53 Pro72 Homozygous Genotype Have a Poorer Survival," *Clin Cancer Res,* vol. 11, pp. 5098-5103, 2005.

[91] O. Erdem, A. Dursun, U. Coşkun, and N. Günel, "The prognostic value of p53 and c-erbB-2 expression, proliferative activity and angiogenesis in node-negative breast carcinoma," *Tumori,* vol. 91, pp. 46-52, 2005.

[92] T. Aas, A. L. Borresen, S. Geisler, B. Smith-Sorensen, H. Johnsen, J. E. Varhaug, L. A. Akslen, and P. E. Lonning, "Specific P53 mutations are associated with de novo resistance to doxorubicin in breast cancer patients," *Nat Med,* vol. 2, pp. 811-4, Jul 1996.

[93] J. C. Bourdon, M. P. Khoury, A. Diot, L. Baker, K. Fernandes, M. Aoubala, P. Quinlan, C. A. Purdie, L. B. Jordan, A. C. Prats, D. P. Lane, and A. M. Thompson, "p53 mutant breast cancer patients expressing p53gamma have as

good a prognosis as wild-type p53 breast cancer patients," *Breast Cancer Res,* vol. 13, p. R7, 2011.

[94] P. Hall and P. Coates, "Assessment of cell proliferation in pathology--what next?," *Histopathology,* vol. 26, pp. 105-112, 1995.

[95] E. de Azambuja, F. Cardoso, G. de Castro, Jr., M. Colozza, M. S. Mano, V. Durbecq, C. Sotiriou, D. Larsimont, M. J. Piccart-Gebhart, and M. Paesmans, "Ki-67 as prognostic marker in early breast cancer: a meta-analysis of published studies involving 12[thinsp]155 patients," *Br J Cancer,* vol. 96, pp. 1504-1513, 2007.

[96] S. Fasanella, E. Leonardi, C. Cantaloni, C. Eccher, I. Bazzanella, D. Aldovini, E. Bragantini, L. Morelli, L. V. Cuorvo, A. Ferro, F. Gasperetti, G. Berlanda, P. Dalla Palma, and M. Barbareschi, "Proliferative activity in human breast cancer: Ki-67 automated evaluation and the influence of different Ki-67 equivalent antibodies," *Diagnostic Pathology,* vol. 6, p. S7, 2011.

[97] E. K. Millar, P. H. Graham, C. M. McNeil, L. Browne, S. A. O'Toole, A. Boulghourjian, J. H. Kearsley, G. Papadatos, G. Delaney, C. Fox, E. Nasser, A. Capp, and R. L. Sutherland, "Prediction of outcome of early ER+ breast cancer is improved using a biomarker panel, which includes Ki-67 and p53," *Br J Cancer. 2011 Jul 12;105(2):272-80. doi: 10.1038/bjc.2011.228. Epub 2011 Jun 28.,* 2011.

[98] G. Falkson, R. S. Gelman, and F. J. Pretorius, "Age as a prognostic factor in recurrent breast cancer," *J Clin Oncol,* vol. 4, pp. 663-71, May 1986.

[99] M. F. Moller, "A scaled conjugate gradient algorithm for fast supervised learning," *Neural Networks,* vol. 6, pp. 525-533, 1993.

[100] Y. LeCun, P. Y. Simard, and B. Pearlmutter, "Automatic Learning Rate Maximization by On-Line Estimation of the Hessian Eigenvectors," in *Advances in Neural Information Processing Systems*. vol. 5: Morgan Kaufmann, 1993, pp. 156-163.

[101] K. Hornik, "Approximation capabilities of multilayer feedforward networks," *Neural Netw.,* vol. 4, pp. 251-257, 1991.

[102] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks,* vol. 2, pp. 359-366, 1989.

[103] M. D. Richard and R. P. Lippmann, "Neural Network Classifiers Estimate Bayesian a posteriori Probabilities," *Neural Computation,* vol. 3, pp. 461-483, 2011/07/19 1991.

[104] P. S. Maclin, J. Dempsey, J. Brooks, and J. Rand, "Using neural networks to diagnose cancer," *Journal of Medical Systems,* vol. 15, pp. 11-19, 1991.

[105] E. Parzen, "On Estimation of a Probability Density Function and Mode," *The Annals of Mathematical Statistics,* vol. 33, pp. 1065-1076, 1962.

[106] B. W. Silverman, *Density estimation for statistics and data analysis*: Chapman & Hall, 1986.

[107] B. Efron, "Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation," *Journal of the American Statistical Association,* vol. 78, pp. 316-331, 1983.

[108] E. R. Dougherty, C. Sima, B. Hua; Hanczar, and U. M. Braga-Neto, "Performance of Error Estimators for Classification," *Current Bioinformatics,* vol. 5, pp. 53-67, 2010.

[109] G. Dreyfus, *Neural Networks: Methodology and Applications* Springer, 2005.

[110] W. A. Yousef, R. F. Wagner, and M. H. Loew, "Comparison of Non-Parametric Methods for Assessing Classifier Performance in Terms of ROC Parameters," in *Proceedings of the 33rd Applied Imagery Pattern Recognition Workshop*: IEEE Computer Society, 2004.

[111] B. Sahiner, C. Heang-Ping, and L. Hadjiiski, "Classifier Performance Estimation Under the Constraint of a Finite Sample Size: Resampling Schemes Applied to Neural Network Classifiers," in *Neural Networks, 2007. IJCNN 2007. International Joint Conference on*, 2007, pp. 1762-1766.

[112] H. Kaizhu, Y. Haiqin, K. Irwin, and M. R. Lyu, "Maximizing sensitivity in medical diagnosis using biased minimax probability Machine," *Biomedical Engineering, IEEE Transactions on,* vol. 53, pp. 821-831, 2006.

[113] H. Demuth and M. Beale, *Neural Network Toolbox for use with MATLAB: User's Guide,* 3 ed. Natick, MA: The Math Works Inc., 1998.

[114] D. Nguyen and B. Widrow, "Improving the learning speed of 2-layer neural networks by choosing initial values of the adaptive weights," in *Neural Networks. IJCNN International Joint Conference on*, 1990, pp. 21-26

[115] V. Velanovich and W. Szymanski, "Lymph node metastasis in breast cancer: Common prognostic markers lack predictive value," *Annals of Surgical Oncology,* vol. 5, pp. 613-619, 1998.

[116] U. M. Braga-Neto and E. R. Dougherty, "Is cross-validation valid for small-sample microarray classification?," *Bioinformatics,* vol. 20, pp. 374-380, 2004.

[117] B. Sahiner, H. P. Chan, and L. Hadjiiski, "Classifier performance prediction for computer-aided diagnosis using a limited dataset," *Med Phys.* , vol. 35, pp. 1559–1570, 2008.

[118] A. K. Jain, R. P. W. Duin, and M. Jianchang, "Statistical pattern recognition: a review," *Pattern Analysis and Machine Intelligence, IEEE Transactions on,* vol. 22, pp. 4-37, 2000.

[119] P. D. Wasserman, *Advanced Methods in Neural Computing*: Van Nostrand Reinhold, 1993.

[120] R. M. Elledge, S. Green, L. Howes, G. M. Clark, M. Berardo, D. C. Allred, R. Pugh, D. Ciocca, P. M. Ravdin, J. O'Sullivan, S. Rivkin, S. Martino, and C. K. Osborne, "bcl-2, p53, and response to tamoxifen in estrogen receptor-positive metastatic breast cancer: a Southwest Oncology Group study," *J Clin Oncol.* , vol. 15, pp. 1916-22, 1997.

[121] K. Maeda, H. Tsuda, Y. Hashiguchi, K. Yamamoto, T. Inoue, O. Ishiko, and S. Ogita, "Relationship between p53 pathway and estrogen receptor status in endometrioid-type endometrial cancer," *Human Pathology,* vol. 33, pp. 386-391, 2002.

[122] E. Y. Cho, Y. L. Choi, S. W. Chae, J. H. Sohn, and G. H. Ahn, "Relationship between p53-associated proteins and estrogen receptor status in ovarian serous neoplasms," *International Journal of Gynecological Cancer,* vol. 16, pp. 1000-1006, 2006.

[123] K. W. Kang, "Mechanism of estrogen receptor loss in breast cancer," University of California, Irvine 2005.

[124] A. Assenza, M. Valle, and M. Verleysen, "A comparative study of various probability density estimation methods for data analysis," *International Journal of Computational Intelligence Systems,* vol. 1, pp. 188–201, 2008.