Interpretability-oriented data-driven modelling of bladder cancer via computational intelligence

Julio Cesar De Alejandro Montalvo

Department of Automatic Control and Systems Engineering

The University of Sheffield

This dissertation is submitted for the degree of Doctor of Philosophy

January 2015

Abstract

The research presented is this thesis entails the study of microarray-based bladder cancer data and the development of new model-based data mining methodologies for accurate prediction of cancer stage, grade and survival. The main focus of the presented research work, from a systems engineering perspective, is on producing models that are more accurate, while maintaining a simple computational structure, interpretable and with good generalisation performance. Such traits deem the developed methodologies as easier to create and use by non-experts.

The presented data-driven computational modelling framework includes a Radial-Basis-Function (RBF) Neural-Fuzzy function, where the universal approximation property is utilised to create an accurate, yet simple, model structure. The scaling-up performance of the developed model is also examined, resulting in a proposal for an enhanced knowledge-capture and model optimisation method. The predictive modelling results show that the RBF-Neural-Fuzzy model outperforms existing modelling attempts in the literature, while identifying clinically relevant gene signatures.

A major contribution of this thesis is the creation of model-based feature-selection framework, as an embedded method for gene signature identification for bladder cancer. For the first time in the literature, an entropy-based iterative algorithm is combined with the previously created RBF model to create an efficient feature selection technique. The Tagaki-Sugeno-Kang (TSK) output layer of the RBF model is used as a feature discriminator, to estimate the relative contribution of each gene to the overall gene signature. The reduced size model (as a result of the iterative feature-selection) achieves more than 80% accuracy on the prediction of patient survival on new

("unseen") patient cohorts, whilst achieving this with less than 25 genes. This is the best performing model-based approach in the literature for this type of cancer, for a gene-signature of less than 25 genes (typical microarray-based signature size in the literature is 100-200 genes).

An in-depth analysis of the generalisation performance of the developed models is carried out by cross-validating distinct microarray data and applying data integration techniques. Three data integration approaches are utilised, to address the well-known issue of data cohort mismatch (for different microarray technologies), and based on the results a model-based non-linear mapping approach is introduced. The obtained results demonstrate how data integration methods for model cross-validation can have a significant increase in the generalisation performance, and enable previously developed models to be used in different patient cohorts.

Acknowledgments

I would like to gratefully and sincerely thank Dr. George Panoutsos, Professor Mahdi Mahfouf and Dr. James Catto for their guidance, understanding, and patience during my graduate studies at The University of Sheffield.

Finally, and most importantly, I would like to thank my family for their support, encouragement and patience.

.

Table of Contents

List of acronyms

(AUC) Area under the Curve

(BN) Bayesian Networks

(BNN) Bayesian Neural Networks

(COG) Centre of gravity

(CI) Computational Intelligence

(CFS) Correlation-Based feature selection

(DNA) Deoxyribonucleic acid

(DOD) Dead of Disease

(FCM) Fuzzy C-means

(FL) Fuzzy Logic

(HDLSS) High dimensional data low sample size data

(HPC) High Performance Computing

(LM) Levenberg-Marquardt

(mRNA) Messenger Ribonucleic acid

(MSE) Mean Square Error

(NN) Neural Networks

(NF) Neural-Fuzzy

(NED) No Evidence of Disease

(PAM) Prediction Analysis for Microarray

(RBF-NN) Radial Basis Function – Neural Network

(RBF-NF) Radial Basis Function Neural-Fuzzy Network

(ROC) Receiver Operating Characteristic

(RMSE) Root Mean Square Error

(SVM) Support Vector Machine

(SVM-RFE) SVM Recursive Feature Elimination

(TSK) Takagi Sugeno Kang

(WFCM) Weighted Fuzzy C-Means

List of figures

List of tables

Chapter 1: Motivation and thesis overview

The healthcare professionals community (medical, biology, chemistry, and engineering) have improved gradually the quality and length of life. It is expected that in developed countries, by year 2020 the female life expectancy will be of around 90 years, however the male life expectancy is not expected to have a considerable life expectancy increase [1]. New vaccinations and novel treatments have been developed for diseases years ago where not possible to treat or cure. These advances in medicine led to a decrease in the mortality rate in many countries. However, it can be said that several diseases are still the focus of research. For example, tobacco-related diseases, which include: respiratory diseases, circulatory diseases and several types of cancer [1].

In healthcare, it is accepted that: "If we live long enough, at certain point of our lives we would develop some type of cancer" [2]. This is why cancer research is of paramount importance, if a cure is found quality and length of life for the patients could be improved. Cancer analysis used to be an area of research destined only for clinicians but as the new technologies emerged and the amount of information increased meteorically, Systems Engineering was required to analyse all of this information. The research presented in this thesis is based on the study of microarray bladder cancer data. Bladder cancer is a highly recursive type of cancer and that it is challenging to treat. The classification of the tumour's behaviour is a crucial point, particularly at the early stage of the cancer when clinicians try to decide which treatment strategy to follow.

This early categorisation of cancer aggressiveness not only helps the patient avoid *unnecessary treatment* but will also allow substantial *cost savings*. The focus of this research is to produce a data-driven computational model that identifies the genes (feature selection) significant to the prediction of stage, grade, and survival of bladder cancer while maintaining *simplicity*, *transparency* and *accuracy*.

1.1 Background and Motivation

Bladder cancer is a type of cancer that is extremely recursive, and depending on the type it can affect the patient's life even after being cured [3]. From a medical point of view, there are two types of cancer according to the evolution of the tumour: aggressive tumours (tumours of poor prognosis and resistant to conventional treatments) and non-aggressive tumours (tumours that respond well to conventional treatment and with good prognosis). However, currently there are no biological markers or reliable parameters to categorise the two types of the disease.

Biology methods based on the analysis of clinical history and biopsies studies are the only routine tools for identification and confirmation of the stage of the disease. From a molecular level since a few years ago genetic profiles and possible markers have been studied. These markers may help us discover the cause and development of the disease; however research is on-going in this area.

Existing analytical techniques, predictive methodologies (regression, prognostic nomograms) and statistical data analysis methods struggle to cope with the inherited noise and uncertainty associated with this type of clinical data therefore yielding average prediction results [4, 5].

In recent years, new techniques have been developed and the study of genetic markers has become more common, however the research is ongoing in this area as at the moment there is still no irrefutable list of genetic markers related to bladder cancer [4-12]. The main prognostic tools are based on histologic stage and grade, and as explained before these tools have their drawbacks. It is logical to assume that there is a research gap, and this is where Systems Engineering may be able to contribute and improve the diagnosis methods via the analysis of both clinical data sets and gene data, in order to find suitable markers that predict cancer stage, grade and survival. The Systems Engineering's aim is to achieve a correct diagnosis, which will lead to the optimisation of the patient's therapy.

For certain cancer types the clinicians perform a number of tests to diagnose the patient. This involves clinical data, chemical tests, and medical examinations and more recently there is interest to investigate gene expression data. Unfortunately, these tests/data are not very well understood. This is where systems engineering and data-driven modelling come in. If hybrid models are built from the test data along with behaviour from cancer biopsies and gene expression data, the understanding of how these tests relate to cancer prediction could be improved, and a treatment therapy could be informed; part of this study is focused on the analysis of microarray data. Microarray is a new technique to analyse tissue samples, and this will be explained in detail in the Chapter 2. Microarray data analysis has opened new possibilities for diagnosis and treatment of numerous diseases, including cancer; however microarray data comes with its limitations. For example, high-dimensionality, low sample size, noisy, missing data and the necessity of applying feature selection methods to identify relevant markers. The diagnosis of numerous malignancies has been improved by the use of microarrays.

For this reason, the search for a robust classifier for the tumours' categorisation and feature selection algorithm has been intensive.

1.2 Research Objectives and Contributions

One of the biggest challenges in bladder cancer prediction is the accurate and early classification; in recent years microarray technologies and related research have helped with this task with feature selection and systems engineering models. Current clinical diagnostic methods are not definitive enough; to date the search for conclusive markers to lead to a precise classification is still ongoing [9, 13, 14].

The main challenges that microarray studies run across are the thousands of genes combined with a small number of samples (patients) and the uncertainty of raw data due to measurement process and variation in the technology. This presents a challenging Systems Engineering classification and identification problem (high dimensionality, low number of samples). To tackle the challenge of high number of features, feature selection algorithms have become indispensable components of the data mining process [15].

The objectives of this research are to:

1. Introduce a Radial-Basis-Function Neural-Fuzzy modelling structure for the analysis of noisy high dimensional low sample size data; the main characteristics of the model are transparency and simplicity.
2. Investigate the scaling-up performance of Radial Basis Function Neural-Fuzzy models using a standard PC and a High Performance Computing (HPC) server. The aim of the research is to find the limit for the maximum number of inputs to use in the model while maintaining low computational complexity and high accuracy.

3. Introduce a new model-based iterative method for feature selection that directly links the relative contribution of each feature to the system's performance.
4. Improve the generalisation performance in microarray bladder cancer data
5. Maintain *simplicity*, *transparency* and *accuracy*.

The main novelty of this research relies on producing models that are accurate, simpler, interpretable, with good generalisation performance (robust) and easier to develop and to be used by non-experts given their simplicity and transparency.

A Neural-Fuzzy algorithm was chosen because it possesses the learning abilities of Neural-Networks, the interpretability of Fuzzy logic and can model non-linearity. Furthermore, Neural-Fuzzy models require less data than Neural-Networks [16]. Apart from the previously mentioned characteristics, Neural-Fuzzy models already proved to make accurate bladder cancer classification [7, 16-19]. Compared to Neural-Networks, Support Vector Machines (SVM) or Logistic Regression, Neural-Fuzzy models deliver comparable or improved accuracy in classification with the advantage of being more interpretable [17]. A drawback of Neural-Fuzzy models is that they encounter problems when the dimensionality is relatively high [20].

In this thesis, the following research contributions have been made:

1. Reduction in the complexity of the model: number of inputs of the model (features) and linguistic statements to describe the model (fuzzy rules).
2. An enhanced rule-base extraction framework is proposed to improve the model's performance for high-dimensional low sample size data (microarray data). With the enhanced rule-base, the scaling-up performance of Radial Basis Function (RBF) Neural-Fuzzy models was improved.

3. For the first time, a Neural-Fuzzy model (Radial-Basis-Function with a TSK output) was applied to microarray bladder cancer data to make a feature selection in the training phase (embedded feature selection): the aim of the iterative feature selection method is to use a measure of uncertainty (fuzzy entropy) to select relevant features during the model-training phase, whilst maintaining the system's simplicity and interpretability and taking into account the interactions between the genes.
4. The inclusion of the cancer stage and grade as extra features of the predictive model is evaluated, thus producing a hybrid gene-clinical data model.
5. Improve the generalisation performance in microarray bladder cancer data: two different data integration approaches were presented for the first time: median adjust and NN mapping of input-output. The results obtained prove that the data integration methods for cross validation of the models helps to have a considerable increase in the performance.

Considering the aforementioned objectives and contributions, and the impact of this disease in the society, the research work described in this thesis is underpinning for the development of new methods of diagnosis and prediction of the behaviour of bladder cancer.

1.2.1 Publications:

Each publication is linked to a Chapter and objective within the thesis and study respectively, for example:

- Chapter 3 is linked to publications 1 and 4 and objectives 1 and 5.
- Chapter 4 is linked to publications 2 and 2b and objective 2.
- Chapter 5 is linked to publications 1, 3 and 4 and objectives 1,3 and 5
- Chapter 6 is linked to objective 4 and 5.

- **Peer reviewed Journal and Conference Publications**

[1] J. De Alejandro Montalvo, G. Panoutsos, M. Mahfouf and J. W. Catto, Radial-Basis-Function Neural-Fuzzy model for microarray signature identification, BIOSIGNALS 2013 - Proceedings of the 4th International Conference on Bioinformatics Models, Methods and Algorithms, Barcelona, Spain 11-14 February, (2013)

[2] J. De Alejandro Montalvo, G. Panoutsos, M. Mahfouf and J. W. Catto, High dimensionality and scaling-up performance of RBF models with application to healthcare informatics, 2014 5th International Conference on Computer and Computational Intelligence, Paris, France, 6-7 December, (2014).

[2b] J. De Alejandro Montalvo, G. Panoutsos, M. Mahfouf and J. W. Catto, High dimensionality and scaling-up performance of RBF models with application to healthcare informatics, International Journal of Machine Intelligence and Computing (IJMLC) (post-conference volume-invited).

- **Invited/Peer reviewed workshops and seminars**

[3] Microarray model-based gene signature identification for the accurate prediction of survival in bladder cancer, University of Sheffield- INSIGNEO Institute for In-silico Medicine Showcase, Sheffield, UK (2014).

[4] Human-Centric Approaches for Modelling Complex Systems', University of Sheffield Engineering Symposium - USES 2013, Sheffield, UK (2013)

1.3 Thesis Outline

The rest of this thesis is organised as follows:

- Chapter 2: definition of cancer, microarrays and a brief literature review is presented. Previous methods used for microarray data analysis, either for feature selection or cancer classification are also covered in this Chapter.
- Chapter 3: This Chapter introduces a Radial-Basis-Function Neural-Fuzzy modelling structure, aiming to maintain simplicity and transparency in the form of a linguistic Fuzzy-Logic rule-base. The proposed methodology is validated by selecting a signature for the identification of the stage, grade and survival of bladder cancer. The signature selection and predictive modelling results are compared to previous research work on the same dataset, showing that the RBF-NF model outperforms the previous modelling attempts by achieving high predictive accuracy (>80% on average).
- Chapter 4: the scaling-up performance of Radial Basis Function Neural-Fuzzy models is investigated. Based on the findings, an enhanced rule-base extraction framework is proposed to improve the model's performance for high-dimensional low sample size data. To overcome the challenges present when high dimensional data is used, a Weighted Fuzzy C-means (WFCM) algorithm for the analysis of high-dimensional low sample size data is introduced. A second contribution of this chapter is a cluster optimisation algorithm based on the Xie-Beni cluster validity index to improve the quality of the initial clusters (rule-base) calculated by the WFCM. Via the proposed framework the scaling-up performance of RBF Neural-Fuzzy models is enhanced, hence the predictive modelling framework can be used without the use of filter-based feature selection methods. The aim is to find the rational limit for the maximum number

of useful inputs (genes) to use in the model while still maintaining low computational complexity and high accuracy.

- Chapter 5: In this chapter, a new model-based iterative method for feature selection based on fuzzy entropy measures is introduced. The presented approach is based on a Radial Basis Function – Neural-Fuzzy which is designed to be equivalent to a Fuzzy Logic TSK-based system. A fuzzy entropy measure is used to directly link the relative contribution of each feature to the system's performance. An iterative algorithm is then used to identify the most relevant features of the process under investigation; the modelling-feature selection is performed in one iterative process. In predicting the patients' survival as a result of their bladder cancer gene signature, the inclusion of the cancer stage and grade as extra features of the predictive model is evaluated, thus producing a hybrid gene-clinical data model. The simulation results confirm that the new approach outperforms existing predictive models in the literature for bladder cancer survival based on gene signature only; the additional novelty of the presented approach relies on the added benefit of producing models that are simpler (considerably less genes in the signature), interpretable, with good generalisation performance and easier to develop and use by non-experts due to the absence of complex pre-processing which is common in this field.
- Chapter 6: In this Chapter, the generalisation performance of the developed models is investigated. The approach studied in this chapter is to cross-validate distinct microarray data by applying data integration techniques. Three different data integration approaches were analysed: quantile discretisation, median adjust and NN mapping of input-output. The latter two approaches are introduced for the first time to a bladder cancer classification algorithm. The results obtained

demonstrate that the data integration methods for cross validation of the models helps to have a significant increase in the performance.

- Chapter 7: Conclusions and future research directions: the conclusions of the thesis and the direction for future research. This covers performance summarisation of the presented work and research proposals towards multi-cohort modelling approaches.

Chapter 2: Microarray technologies and data-driven modelling for cancer

Case Study on Bladder Cancer; definition of cancer, microarrays and a brief literature review is presented. This Chapter also presents previous methods used for microarray data analysis, either for feature selection or cancer classification.

2.1 Cancer Overview

In order to assess the complexity of the problem of cancer and cancer research in the world a brief literature review is presented here. There are more than 200 types of cancer and millions of new cases of cancer are recorded each year [21-24]. According to [23], the most common cancers occurring in the UK are:

- Female breast
- Lung
- Prostate Cancer
- Bowel
- Malignant Melanoma
- Non-Hodgkin Lymphoma
- Bladder

As stated in [23], "More than 331000 people were diagnosed with cancer in the UK in 2011. More than 1 in 3 people in the UK will develop some type of cancer during their lifetime".

For the UK, the survival rates for cancer have increased considerably in the last decades; nevertheless cancer is the cause of more than 25% of all deaths [23].

2.2 Bladder Cancer

Bladder cancer is the 11th most common type of cancer in the world [21, 25]. In the UK, is the 7th most common type of cancer [21, 23].

The bladder is an organ (Figure 2.1) [26] that stores urine and it is located in the abdomen, the majority of the bladder cancer forms in the tissues of the bladder.

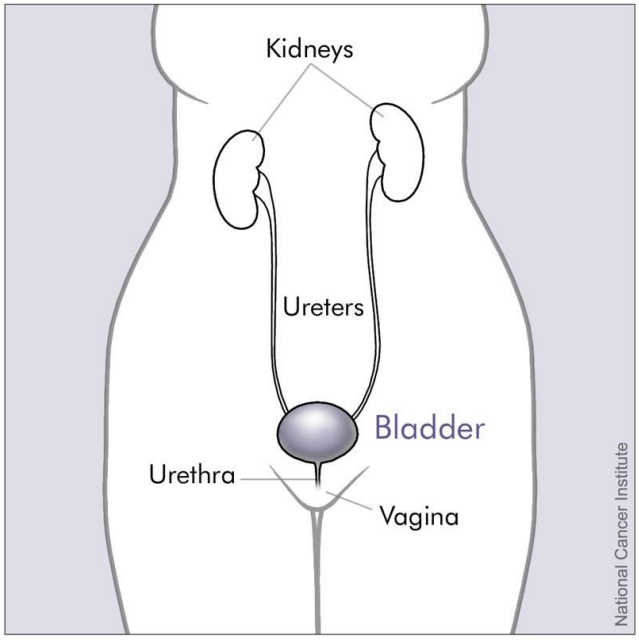

Figure 2.1: Bladder and nearby organs

The risk factors linked to bladder cancer are [27, 28]:

- Age
- Cancer therapies
- Ethnicity

Cancer occurs when something goes wrong with the cell reproduction and the cells do not die but instead they continue reproducing. If those cells are not dying and new cells are reproducing, eventually a tumour may form [23]. The correct classification of future tumour behaviour is one of the biggest challenges in cancer. Whilst it is crucial to avoid unnecessary treatment for indolent tumours, delays in radical intervention for aggressive disease lead to worsening survival and quality of life [29-31]. The prediction of outcome is best performed using pathological stage, grade and various other histological and clinical parameters.

The cancer Stage encoding is based on the staging system that uses numbers to indicate the stage of the cancer, this is defined as follows:

> *"Stage 0a, there is a small area of cancer only in the bladder lining.*
>
> *Stage 0, the cancer cells are confined to the inside layer of the lining of the bladder.*
>
> *Stage 1, the cancer has grown into the layer of connective tissue beneath the bladder lining.*
>
> *Stage 2, the cancer has grown into the muscle of the bladder wall under the connective tissue layer.*
>
> *Stage 3, the cancer has grown through the muscle of the bladder and into the fat layer surrounding it. It may have spread to other organs.*
>
> *Stage 4, the cancer has spread to the wall of the abdomen or pelvis, the lymph nodes or to other parts of the body"* [32].

Similar to the encoding applied to the previous model for the prediction of stage; three grades are used to rate cancer. The Grading of bladder cancer tumours is defined according to:

Grade 1 or low-grade cancer

Grade 2 or moderate/intermediate grade

Grade 3 or high-grade cancer [22].

The risk of disease-progression, as well as the frequent reoccurrences, require extensive clinical monitoring of bladder cancer patients, making this disease one of the most expensive to manage [33]. One of the challenges in the screening of cancer is the search for markers that identify tumour of aggressive and non-aggressive behaviour. This is a crucial point, especially at the early stage of the cancer when clinicians try to decide which treatment strategy to follow. This early categorisation of cancer aggressiveness not only helps the patient avoid unnecessary treatment (often avoiding serious side-effects) but will also allow substantial cost savings.

There is a high lifetime cost on patients with superficial tumours. Removing the bladder can treat the disease, but it may result in various complications that the person has to live with for the rest of their life [28] . Current clinical diagnostic methods are not definitive enough; at this time there are no determinant markers to do a precise detection. Biology methods based on the analysis of clinical history and biopsies studies are the only routine tools for identification and confirmation of the stage of the disease. From a molecular level since a few years ago genetic profiles and possible markers have been studied. These markers may help discover the cause and development of the disease; however research is on-going in this area. Limitations in the accuracy of clinical diagnostic methods have led to the search for more robust biomarkers such as

those derived from gene expression data [7, 34, 35]. In recent years microarray technologies and related research help with the task of making an accurate and early classification of the cancer and with the identification of clinically relevant genes.

2.3 Microarray Analysis

Microarray is a technique to analyse tissue samples. Microarrays make possible the analysis of thousands of genes simultaneously; since thousands of genes are analysed, the data generated from each microarray is enormous. This literature review will focus on giving the basic concepts of microarray and explain what the data represents.

The present section is divided into 3 sub-sections:

1. Microarray Basic concepts.
2. Representation and extraction of information
3. Analysis of Gene Expression Data; different methods for the analysis of microarrays, from statistical to soft computing.

2.3.1 Microarrays Basic Concepts:

Deoxyribonucleic acid (DNA) microarrays are a technology to simultaneously monitor the expression levels for thousands of genes [36]. The process of transcription of genes into messenger Ribonucleic acid (mRNA) and subsequent conversion to form proteins is called Gene expression [36]. DNA microarrays are used to identify disease biomarkers in many applications, for example: in neurological diseases, Alzheimer, multiple sclerosis, diabetes [37, 38]. As shown in Figure 2.2 [39], DNA microarrays are solid supports where gene sequences are immobilised [38].

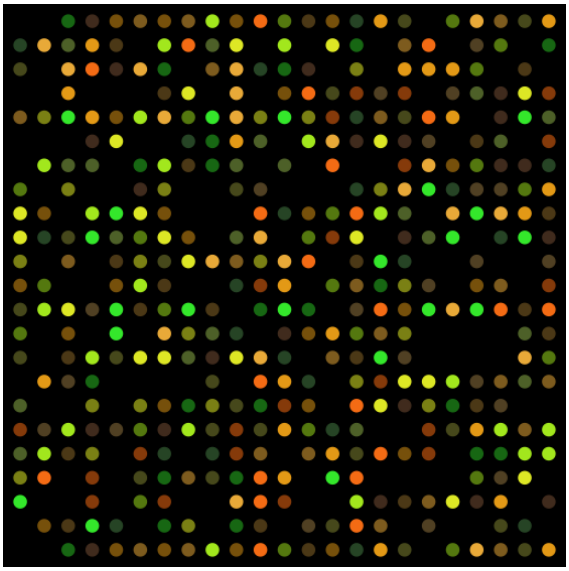

Figure 2.2: DNA microarray

In a microarray, the gene sequences must be attached to their support in a permanent way, since scientists use the position of each spot in the array to detect a gene sequence [38]. The entire process is based on hybridisation probing, defined in [40] as: "a technique that uses fluorescently labelled nucleic acid molecules to identify complementary molecules and sequences that are able to base-pair with one another."

Once the hybridisation is complete, a 'scanner' will examine the microarray. A laser stimulates the fluorescent tags, and the scanner produces a digital image of the array. The image is stored and, as explained in Section 2.3.2, is subsequently analysed.

2.3.2 Representation and Extraction of Information

i. Image Processing and analysis

As mentioned in section 2.3.1, the expression level for each gene can be stored as an image. The processing of the image is the initial step in the analysis of microarray data.

Image processing involves:

- Identification of the spots
- Determination of the area to be analysed
- Assigning the spot intensity [41].

After the image processing and analysis, normalisation is necessary to adjust for any bias that arises from differences in the microarray process [38, 41].

ii. Gene Expression Data Matrices

As stated in [41], there are several representations for the measurements of microarray data:

- Absolute; the expression level of the gene is represented in abstract units.
- Relative; the gene expression of a gene in abstract units is normalised with respect to its expression in a reference.
- Log2; the gene expression values are converted to log2 to eliminate the high variations between the gene's intensities.
- Discrete; the gene expression values are converted to discrete numbers.

Microarray data can be also seen as a vector, where the gene expressions are represented in a vector space [41].

2.4 Feature selection methods applied to microarray

One of the biggest challenges in cancer is the correct classification of future tumour behaviour and it may be achieved via a number of information sources, including clinical and radiological data and potentially, biochemical or molecular tests. However, limitations in the accuracy of these data have led to the search for more robust

biomarkers such as gene expression data. In recent years microarray technologies and related research help with this task. A reliable predictor capable of an accurate assessment at an early stage of the cancer will undoubtedly avoid unnecessary treatment, save costs and in general would improve the patients' quality of life.

Current clinical diagnostic methods are not definitive enough; to date the search for the conclusive markers to lead to a precise classification is still ongoing [9, 13, 14].

The main challenges that these types of studies run across are a) the high dimensionality, translated into tens of thousands of genes combined with a small number of samples (patients) and b) uncertainty of the raw data due to measurement process and variation in the technology. This presents a challenging Systems Engineering classification and identification problem (high dimensionality, low number of samples). To tackle the challenge of high number of features, feature selection algorithms have become indispensable components of the data mining process [15]. The objective of feature selection is to improve the performance of the predictor, provide faster and more computationally inexpensive predictor. There are numerous benefits of feature selection: simplifying data understanding, decreasing the computational complexity, and most importantly decreasing training times [42]. There are three categories for feature selection: filters (typically applied as a pre-processing step), wrappers (optimise a classifier as part of the feature selection procedure) and embedded methods (perform feature selection in the process of training). The goal of this literature review is to offer an analysis of the current feature selection methods applied to microarray data (high dimensional low sample size).

2.4.1 Filter Methods

Filter methods evaluate the correctness of the proposed feature subset by analysing the relation of each gene with the class by the calculation of basic statistics [43, 44]. Filter approaches are the most used feature selection method in microarray literature for gene selection [36, 45].

Filter methods rank the features depending on a score: then, the features with the highest score are chosen and applied as inputs for the classifier [44]. Filter feature selection methods can be separated in two categories: multivariate and univariate. Multivariate methods consider, to some extent, the dependencies between the features; on the contrary, univariate filter methods consider each feature individually [43, 44]. The costs of considering the dependency is being slower, loose some scalability while at the same time still not has interaction with the classifier [24].

i. Univariate Filter Feature selection Methods

The majority of the filter methods belong to the univariate category. The advantages of Univariate filter feature selection methods are that they are fast, scalable and independent from the classifier [24]. The disadvantages of using univariate feature selection methods are that they ignore feature dependencies and lack interaction with the classifier [24]. Some of the most common examples of univariate filter feature selection methods are t-test [46], ANNOVA [46], Information Gain [47].

ii. Multivariate Filter Feature selection Methods

Examples of filter methods applying correlation for feature selection are the Correlation-Based feature selection (CFS) [48] and the fast correlation-based filter method [49].

In [48], the method selects the feature (genes) subsets based on correlation or dependence. The method's objective is to select subsets of genes that show a high correlation with the class but no correlation between the genes. The CFS method reports results (breast cancer) comparable or better than wrapper approaches, with the benefit of being faster.

In [49], the filter method is based on measuring the 'predominant correlation', identifying the features relevant to the class and minimising the redundancy between the selected features. The method is applied to lung cancer, reporting high classification results.

Numerous multivariate filter feature selection approaches [50-53] have been applied to microarray analysis, reporting similar or improved results compared to more specialised methods.

iii. Recent approaches

Recent approaches [54, 55] are considering applying discretisation followed by a filter feature selection algorithm. The authors report an increase in the classification accuracy and the complexity of the model (applied to prostate cancer).

The benefits of using filter techniques are that they can be applied without difficulty to high-dimensional datasets (microarray data); they are computationally inexpensive and fast [24]. A drawback of filter methods is that most techniques consider each feature separately, ignoring feature dependencies, and that they do not interact with the classifier [44]. This could result in an inferior performance of the classifier when compared to more complex methods [44].

2.4.2 Feature Selection: Wrapper Methods

Wrapper feature selection methods optimise a predictor as part of the selection procedure; their computational complexity is high because it grows with the number of features [24]. Consequently, wrapper feature selection approaches have been avoided in recent years. In wrapper feature selection methods, several subsets are produced and assessed [24]. The assessment of each subset is achieved by training and testing each classification model [43]. Wrapper feature selection approaches are popular in machine learning, however due to its large computational cost they are not popular in microarray analysis [43].

Most of the work applying wrapper methods was done in the early years of microarray analysis, and that wrapper methods have not evolve as the same speed as filter or embedded feature selection methods [44]. Despite their high computational cost, several authors [56, 57] state that they have better predictive accuracy.

In [58], the authors evaluate widely applied wrapper feature selection algorithms finding that by using these algorithms the accuracy is improved and the number of genes of the classification model is considerably reduced in size.

In [59], the authors introduce a procedure named successive feature selection. The proposed algorithm consists on separating the genes into subsets (of size s), and subsequently selecting smaller subsets (of size bs) containing the best genes from each subset (of size $bs<s$) based on their classification accuracy. Afterwards, all the selected genes are merged to obtain the ‘top genes subset’.

2.4.3 Embedded Methods

Embedded methods perform feature selection in the process of training the classifier [44]. Similar to wrapper methods, embedded methods also have interaction with the classifier, which increase the computational complexity. However, compared to wrapper methods, the computational complexity is smaller. Embedded methods can be seen as an intermediate solution for feature selection with less computational burden than wrapper methods but higher computational burden than filter methods, without being independent from the classifier [44].

Perhaps the most applied embedded method is a SVM using Recursive Feature Elimination (SVM-RFE) [60]. SVM-RFE is a weighted-based method that trains a SVM with a set of genes and eliminates the genes that are not significant to the solution based on a feature ranking criteria. However, as reported by [44] in their study for breast and cancer prediction, the SVM-RFE achieves comparable or inferior classification accuracy compared to simpler feature selection techniques.

A different SVM approach presented in [61], consists of simultaneously determining a classifier with good classification performance and an small number of features by ‘penalising’ the usefulness of each feature in the elicitation of the model. The approach selects the relevant features according to the width of a Gaussian function, where a small width represents that a feature is important.

2.5 Feature selection methods in high dimensional low sample size data

Although microarray data can be considered as the most representative and complex case of high dimensional data low sample size data (HDLSS), it must not be overlooked that in many different areas HDLSS data is present. Several publications [24, 43, 44, 51, 58, 62-64] have reviewed feature selection algorithms in different areas, for example: image processing, text recognition, financial data, and climate data.

As stated in, [24], the analysis of HDLSS has evolved simultaneously for all the different areas. All the areas come to an agreement that the limitations of the study must be defined: filter feature selection methods are faster but they do not take into account the interactions between the features, wrapper methods consider the interactions but the computational complexity and the necessary time for the calculations augments exponentially, embedded methods suffer from computational high complexity (smaller than wrapper methods but still considerable).

For microarray gene expression feature selection, the *interaction of the features* is of *paramount importance*; moreover a *low complexity is desirable* to work closer with clinicians. Nevertheless, it is essential not to fail to recall that feature selection is half of the necessary work for making a correct classification. Typically, feature selection is done and subsequently a much smaller subset of features is analysed to make the classification. In section 2.6: Machine-learning models for microarray cancer classification, an overview of the most important methods for cancer classifications is presented.

2.6 Machine learning models for microarray Cancer Classification

It is stated previously that the classification of microarray gene expression data is data dependent; furthermore cancer classification is also dependent on the type of cancer. The most common types of cancer analysed are:

- Breast
- Prostate
- Lung

While breast cancer may report high accuracies (circa 90-95 of accuracy) for the prediction of survival using machine learning algorithms, bladder cancer (which is one of the least popular and more recursive) report accuracies approximately 65-80% for the prediction of survival.

2.6.1 Computational Intelligence Modelling for Cancer Classification

Computational Intelligence (CI) can be defined as "the study of adaptive mechanisms to enable or facilitate intelligent behaviour in complex problems" [65]. CI algorithms have proven to be popular in the analysis of microarray data because they can detect complex nonlinear associations between the different variables and offer substantial benefits in terms of tolerance to imprecision and system interpretability. Computational Intelligence includes techniques such as Neural Networks (NN), Fuzzy Logic (FL), Neural-Fuzzy Logic, Support Vector Machine and Bayesian Networks. A review of a number of CI techniques applied to bioinformatics is presented in [66].

In, [7, 17, 18], the authors compared different CI approaches for cancer classification and state that traditional analytic methods fail to give accurate results in microarray data applications because this methods assume biological linearity and use

correlation or dependence to find the relationship between a gene and its class. Within CI there is an area of study called Soft computing. Soft computing could be seen as a number of methods so that real problems could be solved in a similar way as humans solve them [67]. This is one of the most important reasons for the use of Soft Computing, to apply the human reasoning to solve a problem and a human understandable explanation of the model.

Soft Computing includes techniques such as Neural Networks, Fuzzy Logic, Neural-Fuzzy Logic, and Support Vector Machine.

i. Fuzzy Logic

Fuzzy Logic is a linguistic method based on a number of rules that describe the system. The transparency of FL and the possibility of easily interpret the results makes it an attractive and effective method for the analysis of gene expression data [68-71].

An important aspect to take into account at the moment of reducing the number of rules is that in fact is important reduce the rules but the most important is to prove that the reduction of rules does not affects the accuracy of the model. The goal is to have a minimum number of rules with the best accuracy of prediction, not one rule per input.

That is the same case with the number of genes; there is a discussion between the effectiveness of using a large or a small number of genes. As stated previously in this chapter, microarray data is composed of thousands of genes so the main purpose is to find the best genes that could lead us to make a good prediction.

Recent research has shown that a small number of genes are enough for accurate prediction of most cancers, nevertheless the number of genes vary between

diseases [72]. A large set of gene expression will decrease the classification accuracy due to the curse of dimensionality [73]. In this phenomenon, the classification accuracy decreases as the dimensionality increases.

A list with the advantages of Fuzzy Logic method:

- Transparency because of the linguistic rules.
- Easy interpretation of the output because of the Low, Low Medium, Medium, Medium High, High states.
- Rules explaining the model, making easier to clinicians to understand the model.

Due to the characteristics of microarray data (high dimension and low sample size) Fuzzy logic models (as many other methodologies) struggle to make an accurate classification [68].

ii. Neural Networks

Neural Networks are inspired by how the human brain learns and processes information, they have the capability to solve complex tasks [74]. Their concept simulates the behaviour of a biological neural network [74]. While in humans, learning is done by adjusting the synaptic connections between neurons; in NNs, learning is done by adjusting the weights existing between the processing elements of the network [74].

Neural networks can obtain a good performance with higher learning speed in many applications. However, a high complexity of the network (large number of hidden nodes) translates into a slower response of the trained network [75].

A possible disadvantage of neural networks, especially with microarray data, is overtraining. In overtraining, a model can learn a local solution for each example as opposed to finding a global solution [76].

Neural Networks have been successfully applied to the prediction of cancer [77, 78], but some of the informed disadvantages are that the elicited network is hidden within a 'black box', consequently deeming the gain of any insight into the process aspects and into a clinical interpretation [7].

iii. Neural-Fuzzy

The characteristics of Fuzzy Logic and Neural Networks have been discussed in this Chapter; these two methodologies can be combined to form a hybrid Neural-Fuzzy (NF) model. Neural-Fuzzy models combine the learning ability of Neural Networks and the interpreting ability of Fuzzy systems [72]. The fuzzy logic rules of this type of models can be translated into linguistic statements to allow understanding and interrogation of the model.

Neural-Fuzzy systems, are a popular approach for addressing tolerance to imprecision and system simplicity (interpretability) and is widely used in literature [79-82] and more recently also used for the prediction bladder cancer [7, 16-18]. Neural-Fuzzy systems take advantage of the simplicity and tolerance to imprecision of Fuzzy Logic structures and the adaptive learning ability of NN while the inclusion of knowledge to the model is still possible. In general, fuzzy set theory [83] has been extensively applied to pattern classification and FL system have been proven to perform well on uncertain information [84-86]. In terms of their simplicity and interpretability, Neural-Fuzzy models allow model knowledge to be represented in the form of just a

few simple linguistic rules thus rendering such modelling structures appropriate for systems oriented towards human-reasoning (human-centric systems) e.g. clinical decision support systems [87-89].

iv. Support Vector Machines

The support vector machine was initially created to solve classification problems and has been successfully applied to a number of real world problems. Support Vector Machines has exhibited outstanding performance in classification tasks. SVM aims at searching for a hyper plane that separates the two classes of data with largest margin. SVM is shown to be a good classifier for microarray data [90].

Support Vector Machine is a popular method in microarray analysis because it is possible to deal with data with a large number of features and a small number of samples [91]. One of the drawbacks for this method is the high algorithm complexity and the extensive computing requirements of the large-scale quadratic programming tasks. A second problem often mentioned is the poor interpretability as compared to other methods [92, 93].

v. Bayesian Networks

Bayesian networks (BNs) reflect the random nature of gene expression and use Bayes' rule [94]. They are also known as probabilistic networks or probabilistic graphical models. The hypothesis in BN is that gene expression values can be defined by random variables that follow probability distributions [94].

Bayesian networks provide a flexible framework for combining expert knowledge into the modelling process [95, 96]. An additional advantage of BNs is that they are good with modelling the randomness and noise associated with microarray

data [97]. Bayesian networks deal with probabilities but the 'causality' or factors that generated the solution are also important for the network [97].

Bayesian Networks have also been applied to Cancer Prediction [98-101] in particular in the form of a Bayesian Neural Networks (BNN). Bayesian Networks are modelling structures for expressing multidimensional joint probability distributions. The main challenge in using BNN is the necessity to estimate the topology of a BNN from observations, which is not a trivial problem due to the large amount of uncertainty and high computational complexity even for moderate sizes of networks [98, 102, 103].

2.6.2 Machine learning models specific to microarray bladder cancer Stage, Grade and Survival Classification

Specifically, in bladder cancer prediction with microarray, Statistical regression methods (Logistic Regression, Linear Regression) can estimate the progression rate of a population of tumours with limited accuracy (around 70%) [4, 5, 7]. One of the difficulties of statistical methods is that they do not take into account the interaction between the genes; they are only concerned about linear relations between the input and the output.

Specifically to bladder cancer, there are examples in the literature that demonstrate the use of microarray biomarkers (gene signatures) for the prediction of Stage, Grade, Survival, Recurrence and Progression [7, 17, 18, 35, 104-112]. Lauss [113] and Riester [114] demonstrate the use of a SVM to model and predict bladder cancer progression. In [113, 114], the authors identify the most relevant genes for bladder cancer (feature selection) and subsequently develop a model to predict the Stage, Grade, Progression and Survival. In [113] an average prediction accuracy was

reported in the range of 70% to 90%, concluding that signatures with more than 150 genes are needed to obtain robust performance in validation sets.

In [114] it is reported that the simplicity of a predictive modelling structure for bladder cancer survival can be improved by the use of nomograms [115] combined to just 20 genes; however this was achieved at the expense of model accuracy (56% to 75%). Specifically in predicting *bladder cancer progression* the publications [7, 17-19] report interesting results using a Neural-Fuzzy model that aims to be accurate and transparent, and contrary to the study presented in this thesis, the computational simplicity is not essential.

In this Thesis, the use of a Radial Basis Function Neural-Fuzzy (RBF-NF) structure is proposed to address the challenges of: model simplicity, model generalisation and low computational cost. The proposed approach will consist of an embedded method based on a RBF-NF system. The proposed iterative feature selection method takes advantage of a Fuzzy-entropy measure to select relevant features during the model-training phase, whilst maintaining the system's simplicity and interpretability. The biggest strengths of the proposed approach are that the feature selection occurs in the training phase, taking into account the interactions and making it recursive. The proposed approach will be applied to identify suitable gene signatures and predict bladder cancer survival.

2.7 Summary

In this Chapter, the extensive use of microarray in the prediction and treatment of several diseases [9, 105, 107, 113, 114, 116-121] has been discussed. For that reason, there is an increasing amount of data sets available in the public domain; the next logical step would be to validate the results from those experiments. Making these

comparisons may help to obtain more valid and reliable results; however, several difficulties might arise due to the differences in technologies, protocols or analysis used to create each data set.

Several approaches had been made in the past years for analysing microarray data. As described before one of the difficulties that microarray analysis has is the large number of genes, the method has to be effective, fast and as transparent as possible. An important feature of this type of systems is that they are effective working with noisy data. Joined to the trend of low number of genes, researchers are also focused on the robustness of the results and for that reason the cross validation of the results has also become of paramount importance. K-fold validation, Leave-one-out cross validation and Distribution optimally balanced stratified cross-validation are among the most used methods to overcome this issue [122].

As stated in [123],the features in a dataset can be categorised into:

- Relevant: features that help with the classification
- Misleading: features that have a negative effect in the classification
- Irrelevant: features that do not affect (either negatively or positively) the classification
- Redundant: features of a class that has other relevant features.

As stated in [123], "the presence of misleading features will reduce the classification accuracy and, the presence of irrelevant and redundant features will increase the computational burden".

The use of microarrays for cancer classification still represents a great challenge for biologists, clinicians and researchers in general. It must not be forgotten that the amount of information coming from these data is massive and there are still some difficulties when the information is acquired. The biggest challenges to defeat are:

1. There is no standard to make comparable the data obtained from various experiments.
2. The quality of the samples needs to be standardised.
3. Missing values in microarrays
4. Errors and/or noise made in every step of the analysis. From the biologist to the image analysis.
5. The classes are imbalanced

To deal with imbalanced classes bootstrap [124] methods have been used in the literature. Bootstrap refers to resample from the sample data and create an *n* number of 'phantom samples'.

What can be improved in bladder cancer classification is to provide sufficient information and description of any activity in a model, in other words, *transparency* and *simplicity* in the model. In the next chapters, a model with main characteristics of transparency and simplicity will be introduced; this human-centric approach aims to work closer with clinicians in order to identify new combination of genes to predict bladder cancer. This transparency and simplicity can be achieved, at a certain degree, via a RBF Neural-Fuzzy model. Nevertheless, it must never be overlooked that the data modelling performance is at the mercy of the quantity and quality of the measurements of the studied data, in this case the microarray data.

In the next Chapter, a Radial-Basis-Function Neural-Fuzzy modelling structure for the prediction of stage, grade and survival in bladder cancer via microarray data is presented. The resulting model maintains its simplicity and transparency in the form of a linguistic Fuzzy-Logic rule-base. The proposed methodology is validated using a real biomedical case-study, which concerns the signature selection for the identification of the stage, grade and survival of bladder cancer.

Chapter 3: Modelling of microarray gene signatures via Radial Basis Function networks

This Chapter introduces a Radial-Basis-Function Neural-Fuzzy modelling structure for the prediction of stage, grade and survival in bladder cancer via microarray data. The resulting model maintains its simplicity and transparency in the form of a linguistic Fuzzy-Logic rule-base. The proposed methodology is validated using a real biomedical case-study, which concerns the signature selection for the identification of the stage, grade and survival of bladder cancer. The signature selection and predictive modelling results are compared to previous research work on the same dataset, and it is shown that the RBF-NF model outperforms the previous modelling attempts by achieving high predictive accuracy (>80% on average) for a similar-sized gene signature. Crucially, the model is shown to maintain its good performance even when using just 20 genes in the gene based signature.

3.1 Introduction

In the previous chapters the advantages and disadvantages of Neural Networks and Fuzzy Logic have been discussed. It could rather be said that these systems are complementary: Fuzzy logic can deal with inaccurate information on a linguistic level while Neural Networks provide learning and optimisation abilities.

This Chapter describes a Radial Basis Function Neural-Fuzzy Network; this method is a combination of a Radial Basis Function Neural Network and a TSK Fuzzy Model [125].

This combination is possible thanks to their functional equivalence [126]. One of the main characteristics of this type of Neural-Fuzzy is that the output is a linear combination of the inputs; this output is given as polynomial. For the first time, this method is applied to the prediction of bladder cancer stage, grade and survival based on microarray data analysis by means of a low number of inputs (20 as a minimum) and low number of linguistic statements or rules to describe the model (5 rules as a minimum); nevertheless the performance of the model is not sacrificed. As explained in previous chapters, microarray data analysis is challenging because of its dimensionality, complexity and high noise.

An essential characteristic, which is the fundament of the method proposed in this chapter, is to provide enough information and description of any activity in the model. In other words, transparency in the model; and this can be achieved at a certain degree via an RBF Neural-Fuzzy model. Nevertheless, it must never be overlooked that the data modelling performance is at the mercy of the quantity and quality of the measurements of the studied data, in this case the microarray data.

The proposed methodology is based on a systematic algorithmic procedure that aims to pre-process and clean the data, assign initial conditions to the modelling structure and finally iteratively optimise the model. The Radial Basis Function Neural-Fuzzy Network structure addresses the challenges of: a) model simplicity (use of low number of features) b) model generalisation ability (performance in ‘unseen-new’ data and c) low computational cost. The proposed approach is successfully applied to predict bladder cancer survival, stage and grade in three independent data sets.

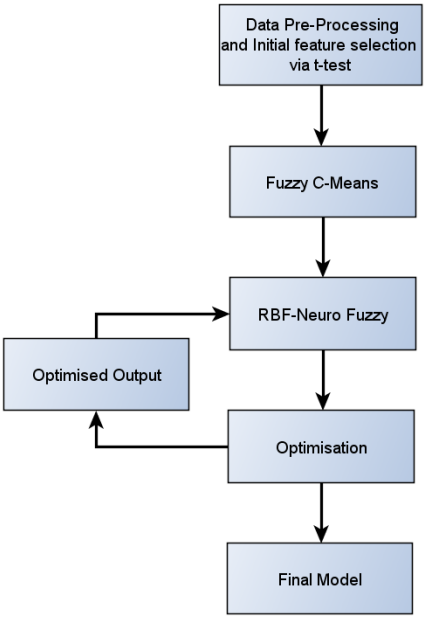

Figure 3.1: Radial Basis Function-Neural-Fuzzy Modelling Structure

The data mining workflow is divided as follows: Data Pre-processing and gene selection with T-Test, Initial Rule-Base Creation via Fuzzy C- Means (FCM), RBF-Based Neural-Fuzzy System, Optimisation of the predictor, Results and Conclusion. The overall approach is presented in Figure 3.1.

The remainder of this chapter is organised as follows: Section 3.2 Data pre-processing and initial gene selection, Section 3.3 Initial rule-base elicitation, Section 3.4 RBF-Neural-Fuzzy Systems, Section 3.5 Levenberg-Marquardt Optimisation Method , Section 3.6 Simulation Results, Section 3.7: Summary.

3.2 Data Pre-processing and Initial Gene Selection

The case-study presented is focused on the prediction of bladder cancer stage, grade and survival using three different bladder data sets: Sanchez-Carbayo [106], Kim [107] and Blaveri [110] (Table 3.1) all of which consist of gene expression data and are considered some of the most complete literature data on bladder cancer gene expression. All the datasets are treated with the same pre-processing procedure.

Table 3.1: Bladder cancer – microarray gene intensity data sets

Data Set	Microarray platform	Number of genes	Number of Samples (patients)
Blaveri	CDNA microarray	10368	80
Sanchez-Carbayo	Affymetrix U133A	22283	90
Kim	Illumina human-6v2.0	43148	165

3.2.1 Normalisation and Missing Values

Prior to any modelling work the data-set is normalised in order to eliminate the high variances between the gene's intensities or the differences in the way that 2 samples are measured [41]. Normalisation is required to remain certain that the differences in two measurements are because different expression values and not because of the different conditions when the measurement was taken [127] . There are numerous normalisation techniques (i.e. local normalisation, normalisation by regression, normalisation by inferring covariates); circa 2003 quantile normalisation gained popularity because it is fast and simple but works equally well than more complex procedures [127]. Quantile normalisation is the process of ordering the values in ascending order in one array, calculate the average between the probes and substitute that intensity with the average and finally change the order to its original [128]. However, no normalisation procedure is flawless; one of the possible drawbacks of quantile normalisation is that the intensities that are greater are forced to fit to the same distribution, decreasing the dissimilarities caused by technical or biological conditions [127].

Quantile normalisation has become a regular procedure [129, 130] to analyse microarrays because it is the default procedure to a very popular software for microarray analysis (Bioconductor [131]). For that reason, the data sets were quantile normalised and then transformed to a log2 scale. The log2 scale allows us to adjust the difference in the intensities to be similar in all the data sets, perhaps it can be said that it

is standard to display the intensities in this scale. The gene expression values were subtracted by the mean intensity to obtain gene-centred log2 values. If the data set had missing values, they were filled using the median, the values with more than 20% missing data were omitted. Missing values are extremely common in this type of data because some of the microarray spots have no expression of the gene at that place or because of errors in the measurement.

To perform the data analysis, Survival outcome is encoded according to Table 3.2.

Table 3.2: Encoding of the Survival Outcome

Survival	Code
No Evidence of Disease (NED)	-1
Dead Of Disease (DOD)	1

The cancer Stage values are 'encoded' into -1 and 1 according to Table 3.3. The Stage encoding is based on the staging system presented in Chapter 2.

Table 3.3: Cancer Stage

Encoded value	Phenotype	Stage
-1	Non-invasive	PTA PT1
1	Muscle invasive	PT2 PT3A PT3B PT4 PT4A

Similar to the encoding applied to the previous model for the prediction of stage; three grades are used to rate cancer and are encoded according to Table 3.4.

Table 3.4: Cancer Grade

Value		Grade
-1	Low/moderate grade	Grade 1 Grade 2
1	High grade	Grade 3

3.2.2 Initial Gene Selection with T-Test

Microarray data set typically have thousands of inputs (genes) and a low number of patient's samples, which added to the previously stated problems, make the classification a challenging task. For that reason, it is necessary to perform an input selection to delete the irrelevant features that are not related to the performance of a classifier. The process of identifying significant features and removing the irrelevant ones is called Feature selection.

The data samples were randomly separated into 'Training' and 'Testing' datasets. The training set is only used to train the model. The testing dataset is only used after the model training is finished in order to test the generalisation performance of the model (i.e. on 'unseen' by the model data), as a form of cross-validation [132].

Training data. The model trains with this data. They have the best performance and represent around 70% of the complete data set.

Testing data. After the model is trained the training data makes predictions on the testing data, represents around 30% of the complete data set. This is the most important parameter to review.

After the pre-processing of the data, the student's distribution t-test is used as an initial feature selection gene based on the p-values.

This is a common pre-processing step in microarray gene selection [128, 133], aiming at removing the irrelevant – 'easily identifiable' – to the process genes. The t-test is a statistical test used to test premises regarding a population. Essentially, a level of significance (what the p-value will be compared to) was selected in order to determine how likely the hypothesis being tested may occur purely by chance[134]. The

disadvantage of using the t-test feature selection method is that it ignores feature dependencies and lacks of interaction with the classifier.

A selection of the genes is made, individually, for the data sets of: Sanchez-Carbayo, Blaveri and Kim, using the top 20 genes for the prediction of survival and 150 genes for the prediction offstage and grade, as selected with the t-test.

3.3 Initial rule-base elicitation via Fuzzy C-Means

To 'translate' the raw datasets into knowledge a Fuzzy C-means algorithm was applied for the elicitation of the initial rule-base. The FCM method [135, 136] is frequently used in pattern recognition; the main justification for using it at this point is because the resulting clusters-rules can be used directly in the form of an RBF model thus simplifying the model creation process, as shown in [137, 138]. This rule-base is then 'translated' into a Radial-Basis-Function Neural-Fuzzy structure, and is finally parametrically optimised via the Levenberg-Marquardt function-minimisation algorithm. The essence of FCM is the exemplification of the similarity that a point shares with each cluster (rule); this exemplification is made with a function (membership function). FCM is based on the following objective function:

$$J_{m=} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \left\| x_i - c_j \right\|^2, 1 \le m < \infty \qquad \textbf{(3.1)}$$

where m is any real number greater than 1, u_{ij} is the membership degree of x_i in the cluster j, x_i is the measured data, c_j is the centre of the cluster. The membership u_{ij} and the cluster centres c_j are calculated by:

$$u_{ij} = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i - c_j \right\|}{\left\| x_i - c_k \right\|} \right)^{\frac{2}{m-1}}} \quad , \quad c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} * x^j}{\sum_{i=1}^{N} u_{ij}^{m}} \qquad \textbf{(3.2)}$$

Each sample will have a membership in every cluster; a higher membership would translate into a higher degree of similarity between the sample and the cluster. Each derived information granule (data-cluster) depicts a process rule in the Fuzzy Logic domain. Figure 3.2 illustrates the 'information granules' divided into three steps: raw data; 1) each data point is considered into the input space, 2) input space granulation via FCM; the initial clusters (information granules) are produced via FCM, 3) Neuro-Fuzzy Rules; the third sub figure shows the initial values for the membership function after the granulation, these values are going to be optimised later.

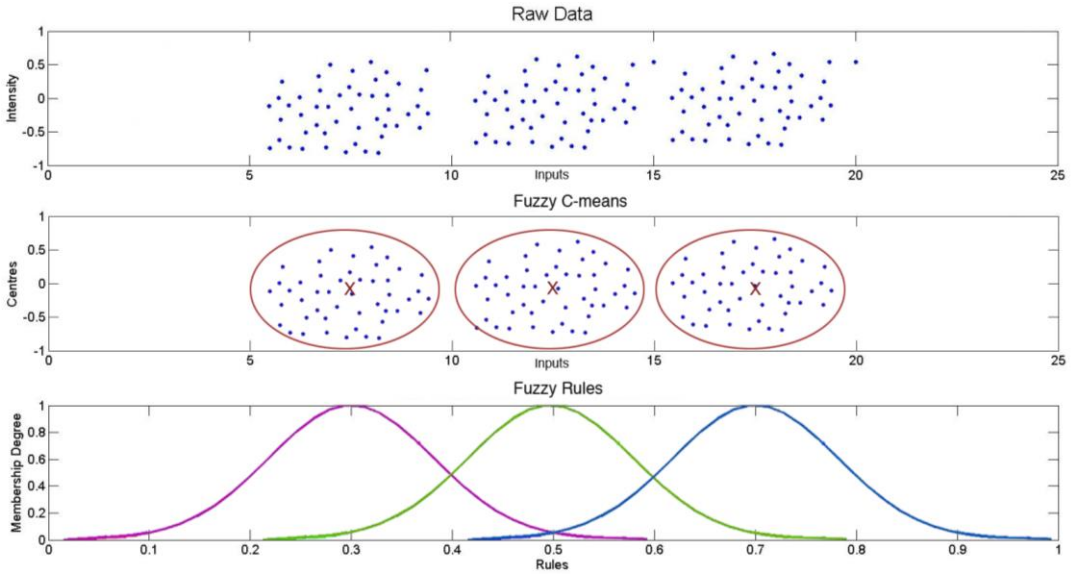

Figure 3.2: Data clustering towards 'information granules' in the Fuzzy Logic domain

The data to be introduced is composed by all the patients and genes plus the real stage, grade and survival outcome. The output from Fuzzy C-Means contains the centres of the cluster, sigma and weights of the outputs. In this approach a threshold for the sigma (width of the membership function) is applied, this will help to make the rules more general, i.e. less specific to the training data, improving the performance in the testing data set.

Another characteristic of Fuzzy C-Means is that the number of clusters (rules) needs to be specified. As explained in the introduction, the proposed model would be

computationally simple and with the lowest number of inputs. To ensure the low computational complexity 5, 10 and clusters (rules) are selected as the initial number of clusters.

An interesting remark is that there is a constraint in the number of genes that can be applied to FCM, if more than a thousand (1000) genes are used, the centres of the clusters became the same among all the samples and only just vary between genes. This means that the clustering thought that the centre was the same for all the samples; it gave a centre for every gene. The next step of the process is the Radial Basis Function calculation.

3.4 RBF- Neural-Fuzzy System

Microarray datasets pose a significant data-mining challenge because of the associated high dimensionality, low number of samples, as well as complexity, non-linearity and high noise (uncertainty). A Neural-Fuzzy system is basically a system that represents information in an interpretable approach but also have the learning ability of a Neural Network, reducing the disadvantages showed by both methods when they are applied by themselves. An RBF-Neural-Fuzzy system offers a good balance of performance and simplicity while being tolerant to some imprecision and being capable of accurate model representations even when few samples are available [87, 88]. In addition, the Fuzzy Logic rule-base ('model knowledge') can be easily interpreted by clinicians as this is in the form of simple linguistic sentences (IF-THEN rules).

The linguistic statements are given in the form:

IF Gene 1 is x intensity and Gene 2 is y intensity **THEN** Output is Z.

The intensities of the Genes are divided in 7 categories according to their value, as in Table 3.5:

Table 3.5: Interpretation of the Normalised Gene Intensity Range

Gene Intensity	Range
Very Low	-1.00 to -0.72
Low	-0.71 to -0.44
Low Medium	-0.43 to -0.16
Medium	-0.15 to 0.12
Medium High	0.13 to 0.4
High	0.5 to 0.68
Very High	0.69 to 1.00

If a Fuzzy logic system is considered. The consequent part of the linguistic statement (…THEN Output is) can be:

a) Fuzzy Set (Mamdani rule-base); output is given as a membership function

b) Singleton (Mamdani singleton); output is given as a single point

c) Lineal Function (Takagi-Sugeno-Kang, TSK); output is given as a polynomial

The main justification to choose TSK type of Neural-Fuzzy is because the output of this system is a linear combination of the inputs; this output is given as a polynomial. In this fashion, it can be analysed how each individual input behaves in the system, allowing the clinicians to interrogate the model. In the Chapter 5 (Chapter 5: A new RBF-NF entropy approach for model-based input selection) a model that analyses the behaviour of a gene in the model and based on that, identifies if that gene may or may not be significant for the classification stage as it is not ‘involved’ in the training of a particular linguistic rule in the rule-base is presented. A new feature selection could be generated; this new feature selection approach would take into account the interactions between the genes.

The method proposed in this Chapter is based on Fuzzy Logic systems having the centre of gravity (COG) defuzzification, the product inference rule and a TSK fuzzy output space, which can be expressed as follows [139]:

$$y = \sum_{i=1}^{p} Z_i \left[\frac{\prod_{j=1}^{m} \mu_{ij}(x_j)}{\sum_{i=1}^{p} \prod_{j=1}^{m} \mu_{ij}(x_j)} \right] \quad \textbf{(3.3)}$$

where $\mu_{ij}(x_j)$ is the RBF function of x_j that belongs to the i-th rule:

$$\mu_{ij}(x_j) = e^{-\left(\frac{(x_j - c_{ij})^2}{\sigma_{ij}^2}\right)} \quad \textbf{(3.4)}$$

where c_{ij} and σ_{ij} are the centre and the width of each membership function, respectively, *m* the number of inputs and p the number of rules. Equation 3.3 can be re-written as follows:

$$y = \frac{\sum_{i=1}^{p} z_i m_i(x)}{\sum_{i=1}^{p} m_i(x)} \quad (3.5)$$

where $m_i(x) = e^{(-\|x - c_i\|^2 / \sigma_i^2)}$ is the degree of membership of the current input vector *x* to the i-the fuzzy rule. Finally, using the radial basis function (RBF) definition:

$$g_i(x) = \frac{m_i(x)}{\sum_{i=1}^{p} m_i(x)} \quad \textbf{(3.6)}$$

The neural-fuzzy input–output relationship then becomes:

$$y = \sum_{i=1}^{p} z_i \, g_i(x) \quad \textbf{(3.7)}$$

$$z_i = a_1 x_1 + a_2 x_2 + a_3 x_3 + \cdots a_j x_j \quad \textbf{(3.8)}$$

Figure 3.3 shows the equivalent NN structure of the RBF model, where the input, rule-base (hidden layer) and output layers can be identified. The input layer is composed by the number of inputs of the system; the middle layer of the RBF is

calculated with the sigmas, output weights and centres of the membership function. Only one output is obtained, this process is repeated for q (number of samples) times.

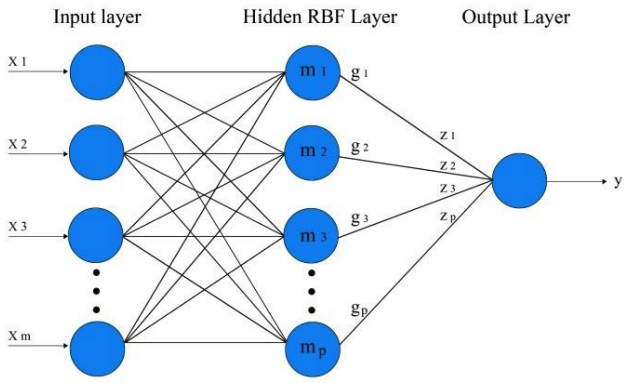

Figure 3.3: Radial Basis Function Neural-Fuzzy Structure

RBF-NF have been used for different applications in microarray analysis, including: breast cancer classification [117, 140], cancer classification in colon and leukaemia [141], multiple sclerosis [119], and lung cancer [142].

Every aspect of the a data-driven modelling approach is important and in the analysis process it was discovered how normalisation affects the data, the number of inputs a method can work with (complexity dependant). Taking these challenges into account, The Levenberg-Marquardt (LM) algorithm was applied for the optimisation [143]. The developed model was parametrically optimised via a suitable function minimisation algorithm. In this approach, the Root Mean Square Error (RMSE) between the training data and the model predicted data was used as the cost function to be minimised. The RMSE is defined as:

$$RMSE(\emptyset) = \sqrt{MSE(\emptyset)} = \sqrt{\frac{\sum(\widehat{\emptyset}-\emptyset)^2}{n}} \qquad \textbf{(3.9)}$$

where $\widehat{\emptyset}$, is the Real Stage, grade or survival of cancer, $\emptyset$ the Predicted value and n the number of elements.

3.5 Levenberg Marquardt Optimisation

The Levenberg-Marquardt [144, 145] algorithm is a standard technique to solve non-linear least-squares problems [146], with this optimiser it is possible to go from a modelling structure to and optimised predictor. In this approach, the RMSE between the training data and the model predicted data is used as the cost function to be minimised.

The algorithm is a combination of the steepest (gradient) descent and the Gauss-Newton method, depending on how far from the solution the method is. The assumption is that if the error is increasing, steepest descent should be used, if the error is decreasing, the algorithm should gradually shift to Gauss Newton.

The Levenberg-Marquardt algorithm is described by the following equation:

$$w_{i+1} = w_i - (H + l\, diag|H|)^{-1} d \qquad \textbf{(3.10)}$$

The equation is a variation of the deepest descent equation, adding the H, l and d. H is an approximation to the Hessian, which is obtained by averaging outer products of the first order derivative (gradient). The derivative is expressed by d and l is the blending factor that determines the mix between steepest descent and the quadratic approximation.

It is important to emphasise that the optimisation approach is a standard technique; the Levenberg-Marquardt algorithm is beyond the scope of this Chapter. In this approach, the total variables to optimise for the optimisation are:

- Centres of inputs
- Sigma of inputs
- Output weights (TSK polynomial)

The minimum value of sigma is set to 0.3; this approach would allow us to exploit the generalisation abilities of a Neural-Fuzzy model. The wider the rules are, the more general the model, increasing the performance to the unknown data (testing).

Overall, the presented data-mining workflow provides an efficient and fast method for capturing numerical data-based information and converting it to a linguistic knowledge-base with a predictive capability.

3.6 Simulation Results

This section is sub-divided into three different parts as follows:

A. Simulation results for Survival: a model is produced and validated using the previously mentioned data sets for the prediction of survival of bladder cancer.

B. Simulation results for Stage and Grade: the model is validated using a real biomedical case-study, which concerns the prediction of the stage and grade of bladder cancer.

C. Comparison to existing literature results: the obtained results for the prediction of stage, grade and survival are compared to previously published results.

D. Fuzzy Logic-type linguistic rule-base: an example of the fuzzy rule-base describing the behaviour of the model.

This section is focused on the prediction of Cancer Survival, stage and grade; the main focus is to identify the best possible combination of rules and training iterations for their prediction.

The number of inputs used in this study is equal to the results for comparison of the models; Lauss[113] (150 inputs stage and grade) and Riester [114] (20 inputs

survival). In this study, the RBF Neural-fuzzy model is applied to the Sanchez-Carbayo, Blaveri and Kim data set to predict Survival rate. At the same time the number of rules and iterations to train the model is assessed to identify how the performance is affected.

As mentioned earlier, there are several unknown parameters that should be taken into account for the prediction of survival; the number of rules for the model, the number of iterations for the trained model. The following plan for the prediction of Survival rate in the Sanchez-Carbayo, Blaveri and Kim data sets was produced, the intention was to be systematic with the computational time required. The methodology applied to this study is explained below (Figure 3.4):

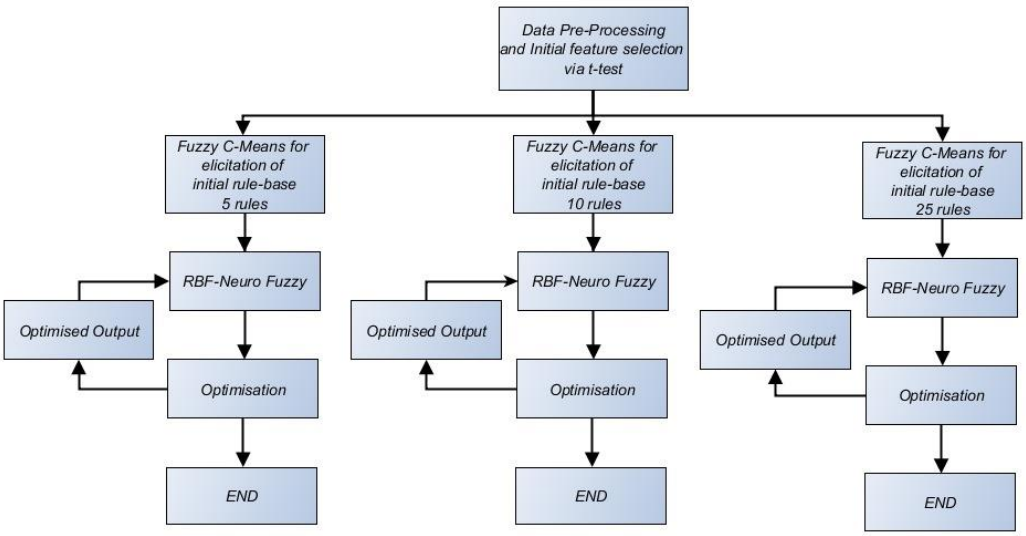

Figure 3.4: Modelling structure for the prediction of survival in bladder cancer.

Train the model with; 5, 10 and 25. Train the model with the selected genes:

- 20 (Survival)
    - 10-50 iterations
    - Cross validation (10 models)

The Survival rate was chosen because it is one of the most complex parameters to predict, the Training structure (number of rules) obtained from the modelling with Survival was later tested with the prediction of Stage and Grade.

3.6.1 Survival Prediction

The RBF-NF model was developed using microarray data intensities only for patients with Muscle-Invasive Cancer, this is to make a fair comparison with the previously published results from Riester [114].

The classification functions of Specificity, Sensitivity and Accuracy are used as measures of performance [147]. The data samples were randomly separated into 'training' (70% of the patients) and 'testing' (30 % of the patients) data-sets. The training set is only used to train the model, and the testing data-set is only used after the model training is finished in order to test the generalisation performance of the model (i.e. on 'unseen' by the model data), as a form of cross-validation. The model was trained with 20 inputs, 5, 10 and 25 rules and cross-validated 10 times. Survival was encoded according to Table 3.3. In order to select the best model it is relevant to analyse the behaviour of each individual model. From a modelling perception, the best model should be a combination of the best performance, lowest computational complexity and practical. The first model to analyse is Sanchez-Carbayo.

a) Sanchez-Carbayo Data Set

The results shown in Table 3.6 are the median of the 10 models for Accuracy, Specificity and Sensitivity respectively, the standard deviation is also shown to have an awareness of how large is the deviation between the values.

Table 3.6: Sanchez-Carbayo performance for Survival

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
5 rules	**Performance (%)**	98	100	97	82	82	82
	Standard Deviation	1	0	3	7	10	14
10 rules	**Performance (%)**	99	100	99	80	75	86
	Standard Deviation	7	0	1	5	7	9
25 rules	**Performance (%)**	99	100	98	84	80	87
	Standard Deviation	1	0	2	5	13	8

5 rules

The testing accuracy performance had a variation from 81 to 84% (Table 3.6), showing that with a higher number of iterations the performance for testing decreases, having the point where the Accuracy, sensitivity and sensibility were more balanced at 30 iterations. An interesting remark is that because of the low complexity of the number of rules and inputs, this model is trained and 10 fold cross validated in less than 60 minutes.

10 rules

The performance of the model when the number of rules was increased did not affect the performance; in fact they were similar to the results obtained with 5 rules.

On the other hand, the complexity of the model did increase the modelling time to the double (two hours) but because the number of inputs used (twenty) is not high, it is still practical. Compared to the results obtained with five rules (Table 3.6), the results with ten rules do not shown an improvement in the performance or the similar balance.

25 rules

As shown in Table 3.7, with twenty five rules the performance of the model was slightly better but less balanced if it is compared to the previous results with less number of rules. Nevertheless, the complexity and the increase in training time that the increase of rules brought are not reasonable.

b) Blaveri Data Set

The Blaveri data set is the smallest in number of samples (patients) making the computational complexity even lower, these models was trained and 10 fold cross validated in less than 50 minutes.

Table 3.7: Blaveri performance for Survival

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
5 rules	**Performance (%)**	100	100	100	90	90	90
	Standard Deviation	0	0	0	6	16	5
10 rules	**Performance (%)**	100	100	100	91	96	90
	Standard Deviation	0	0	0	3	10	2
25 rules	**Performance (%)**	100	100	100	89	86	90
	Standard Deviation	0	0	0	3	17	2

5 rules

The results shown in Table 3.7 reflect a similar behaviour that the one shown for five rules in the Sanchez-Carbayo data set, at 50 iterations the model is more balanced.

10 rules

Similar to the behaviour with the Sanchez-Carbayo data set, the added complexity of the model increases the modelling time to the double but compared to the results obtained with five rules (Table 3.7), the results with ten rules do not shown a significant improvement in the performance or the balance.

25 rules

With twenty five rules the performance of the model was not better if it is compared it to the previous results with less number of rules. With the results of performance and the increase in complexity and in training time is not reasonable to select this model as the top one.

c) Kim Data Set

Table 3.8: Kim performance for Survival

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
5 rules	**Performance (%)**	97	95	99	67	64	70
	Standard Deviation	2	4	1	10	23	20
10 rules	**Performance (%)**	98	98	99	65	57	70
	Standard Deviation	1	2	1	10	14	11
25 rules	**Performance (%)**	96	93	97	68	63	74
	Standard Deviation	3	6	2	8	2	2

5 rules

The results shown in Table 3.8 reflect a dissimilar behaviour that the one shown for five rules for the Sanchez-Carbayo and Blaveri data sets. The Kim data set is the largest in number of samples (patients) making the computational complexity higher. Another important remark is that this data set had more missing values than any other

from the previous results shown. An interesting remark is that the standard deviation between the ten models is high.

10 rules

The performance of the model when the number of rules was increased did not affect the performance in a positive manner; in fact they were similar to the results obtained with 5 rules. Compared to the results obtained with five rules (Table 3.8), the results with ten rules do not shown an improvement in the performance or the balance.

25 rules

As shown in Table 3.8, with twenty five rules the performance of the model was not better or balanced if it is compared to the previous results with less number of rules.

The analysis of the results for the prediction of survival reveals that is possible to generate a simple model with five rules, reducing the complexity and training time. The generation of a model with more than five rules is not justified; the results revealed a number of disadvantages from the increase of the number of rules, such as; not significant improvement in the performance, increase in the training time, unnecessary computational complexity. The Stage and Grade will be done with five rules and from 10-50 iterations.

The computational complexity of the models would be superior since one hundred and fifty inputs are going to be used.

3.6.2 Stage and Grade Prediction

From the previous results for the prediction of survival, it can be concluded that the elicitation of a model using five rules is adequate to obtain comparable or improved performances to the ones obtained with a higher number of rules (ten or twenty five).

The model will also benefit from the reduced of the number of rules by reducing the training iterations, making the model simpler. Applying the same methodology to the prediction of stage and grade for Sanchez-Carbayo, Blaveri and Kim tested the proposed hypothesis and the results were similar to the obtained with Survival. Stage and grade were encoded according to Table 3.3 and 3.4, respectively. The results presented in this section correspond to a model with 5 rules and 150 inputs.

I. Stage

Table 3.9: Performance for Stage

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	99	99	100	94	97	92
	Standard Deviation	.05	1	0	2	4	3
Blaveri	**Performance (%)**	100	100	100	60	75	57
	Standard Deviation	0	0	0	8	22	9
Kim	**Performance (%)**	88	90	81	70	79	56
	Standard Deviation	5	9	12	6	16	19

a) Sanchez-Carbayo

For Sanchez-Carbayo, the results showed in Table 3.9 demonstrate that the increase in the number of inputs did not decrease the performance or increase the iterations for the model to be trained. Similar to the results for survival, the best performance was found at 10 iterations.

b) Blaveri

Differing to the results found for survival, the results for the prediction of stage in the Blaveri data set are less than average. This is an interesting remark; perhaps the complexity of the increase in the number of inputs affected the model.

Nevertheless, it would be highly unpractical to produce a model with a higher number of rules because the computational complexity would increase exponentially. With 10 iterations, the best performance for Accuracy, sensitivity and specificity is shown in Table 3.9.

c) Kim

As explained in the previous section, Kim's data set is the largest in the number of samples. There was no surprises in the performance of the model for the prediction of stage, it was similar to the one obtained for the prediction of survival. The best performance is shown in Table 3.9.

II. Grade

Table 3.10: Performance for Grade

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	99	98	99	94	97	92
	Standard Deviation	1	2	1	3	4	3
Blaveri	**Performance (%)**	99	99	100	97	76	98
	Standard Deviation	6	2	0	2	25	1
Kim	**Performance (%)**	91	93	89	80	83	74
	Standard Deviation	4	4	10	4	7	18

a) Sanchez-Carbayo

For Sanchez-Carbayo, the results showed in Table 3.10, demonstrate the constant performance for all the prediction models (survival, stage and grade) produced. The increase in the number of inputs did not decrease the performance or increase the iterations for the model to be trained. Similar to the results for survival and stage, the best performance was found at 10 iterations.

b) Blaveri

Differing to the results found for stage and similar to the results for the prediction of survival stage, the results for the prediction of grade in the Blaveri data set are high. With 10 iterations, the best performance for Accuracy, sensitivity and specificity is shown in Table 3.10.

c) Kim

The results for the prediction of grade in Kim's data set are the highest produced by all the predictor models in this study. The best performance was obtained at 30 iterations; it is shown in Table 3.10.

3.6.3 Fuzzy Logic-type linguistic rule-base

The models presented in this Chapter maintain a transparent Fuzzy Logic-type linguistic rule-base. Figure 3.5 shows a sample of the rule-base describing the behaviour of the model. For simplicity, just two rules are shown (one for 'negative outcome' and one for 'positive outcome'); these are shown for five out of the 20 genes in the gene signature.

Two of the linguistic IF-THEN rules that describe the model are shown below to demonstrate the transparency (interpretability) of the modelling method. The corresponding numerical values of the linguistic hedges 'high', 'medium' etc. are determined by the optimisation algorithm via the training data-set. The equivalent linguistic-numerical interpretation of the normalised gene intensity is shown in Table 3.5.

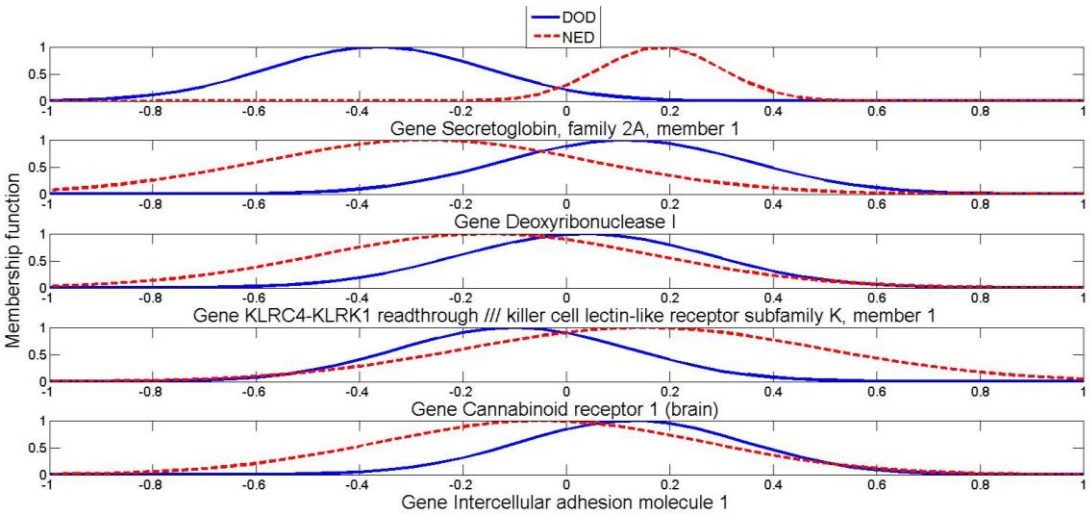

Figure 3.5: Example of a Radial Basis Function-Neural-Fuzzy rule base, here for simplicity just two rules are shown.

Rule 5 (DOD):

IF the intensity of

the Gene 'Secretoglobin, family 2A, member 1' is Low Medium and

the Gene 'Deoxyribonuclease I' is Medium and

the Gene 'KLRC4-KLRK1 readthrough/// killer cell lectin-like receptor subfamily K, member 1' is Medium and

the Gene 'Cannabinoid receptor 1 (brain)' is Medium and

the Gene 'Intercellular adhesion molecule I' is Medium High

THEN the Patient will decease as results of the disease

3.6.4 Comparative Study

a) Survival Outcome

Table 3.11 shows the performance of the RBF Neural-Fuzzy model compared to previous results published by Riester [114]. The Riester study makes use of three

independent datasets (Sanchez-Carbayo [106], Blaveri [110] and Kim [107]) to develop a hybrid model using both SVM and a clinical nomogram [115] to assist with the predictions based on 20 inputs. The RBF-NF model exhibits a better balanced performance (Area under the Curve of the Receiver operating characteristic curve) in two of the three cohorts (other performance indicators were not published). It is important to note that the RBF-NF model achieves a superior or performance in the Sanchez-Carbayo and Blaveri case.

The simplicity of the RBF-NF modelling structure could be crucial for developing easy to use clinical advisory tools. For example, the NF-based structure allows the direct interpretation of the system's rule base to natural language (via Fuzzy Logic linguistic statements – see Figure 3.5), which can aid the development of human-centric systems for use in healthcare.

Table 3.11:Performance of Survival using microarray data. For comparison purposed the results in this example are shown as the area under the curve (AUC) of a ROC plot

	Survival	
	Riester [114] (SVM + Nomogram 20 genes)	**RBF Neural-Fuzzy 20 Inputs**
Sanchez-Carbayo	0.74	0.82
Blaveri	0.76	0.90
Kim	0.75	0.67

b) Stage and Grade Outcome

Tables 3.12 and 3.13 show the performance obtained from prediction of Stage and Grade. The presented model obtained better or comparable performances to previously published results [113] with a SVM approach using 150 genes. Table 3.12 shows a comparison between a SVM model with 150 inputs and the RBF model with 150 inputs. The RBF model performed better for the Sanchez-Carbayo data set but for

Blaveri Lauss had a better performance. No results for Kim were found to make a comparison.

Table 3.12: Comparison of results from the prediction of Stage to existing publications in the literature

Stage (Accuracy)		
	Lauss (SVM-150 genes) [113]	**RBF Neural-Fuzzy (150 genes)**
Sanchez-Carbayo	87 %	94 %
Blaveri	85 %	60 %
Kim	-	70 %

Table 3.13 shows a comparison between the same Lauss SVM model with 150 genes and the RBF Neural-Fuzzy model with 150 genes. The RBF model performed better for the Blaveri data set and or Sanchez-Carbayo. No results for Kim were found to make a comparison.

Table 3.13: Comparison of results from the prediction of Grade to existing publications in the literature (Accuracy)

Grade		
	Lauss (SVM-150 genes) [113]	**RBF Neural-Fuzzy (150 genes)**
Sanchez-Carbayo	80 %	94 %
Blaveri	86 %	97 %
Kim	-	80 %

3.7 Summary

The study of Cancer is of great significance due to several factors; including the increasing mortality rate, to help avoid unnecessary treatment and from a Bioinformatics perspective to help clinicians to understand these studies. There are several methods used in present days that do not take in account the subtle relation between the genes and the complexity of Gene Expression data, for that reason it is relevant to investigate this subject. The main problem that these types of studies run

across is the high dimensionality, translated in thousands of genes but a small number of samples.

As there are no equations that represent the behaviour of the genes, a predictor model must be produced. With high dimensionality also comes noise in the intensities, a large presence of irrelevant and redundant genes. The goal of these studies is to produce a model capable of making a prediction based on the existing data that is efficient, could be understood by clinicians (transparent) and with the lowest computational cost. Joined to the above description there is another quality that the predictor must give, the selected genes must be relevant from a clinical point of view. For that reason medical and engineering expertise must work together, to validate the performance of the study.

The proposed RBF-NF methodology has successfully been applied to the case study of bladder cancer prediction with respect to the patient's stage, grade and survival. A list with the advantages of this method:

- Transparency because of the linguistic rules.
- Easy interpretation of the output because of the Very Low, Low, Low Medium, Medium, Medium High, High, Very High states.
- Minimum number of rules explaining the model, making easier to clinicians to comprehend the model.
- Universal approximation ability (RBF)
- Low computational cost

Compared to previous modelling attempts from Martin Lauss [113] and Riester [114] based on SVM, the RBF-NF method shows improved performance in the same datasets. However, the attractiveness of this method is on the transparency that the rule-base exhibits and the good generalisation performance (even with just 20 genes and 5 rules) as compared to previous modelling attempts on the same dataset. The rule-base's transparency and interpretability, can aid the clinicians to directly interrogate the resulting model (human-centric system) and examine how the model uses individual genes and their intensity to provide predictions on the stage, grade and survival of bladder cancer.

Chapter's summary of achievements:

- Development of a Radial-Basis-Function Neural-Fuzzy Linguistic Modelling algorithm (from data clustering to optimisation)
- An RBF-NF model was applied for the accurate prediction of stage, grade and survival of bladder cancer.
- The predictive modelling results show that the RBF-NF model outperforms the previous modelling attempts by achieving high predictive accuracy (>80%).
- The model is shown to maintain its good performance even when using just 20 genes in the gene based signature.

The achievements summarised above are linked to one conference publication (Biostec 2013), and The University of Sheffield Engineering Symposium - USES 2013, Sheffield, UK (2013).

On the next chapter, the scaling-up performance of Radial Basis Function Neural-Fuzzy models is investigated. The aim is to find the rational limit for the maximum

number of useful inputs (genes) to use in the model while still maintaining low computational complexity and high accuracy. Nevertheless, it must not be overlooked that are several challenges to defeat:

- the computational complexity of this models will increase exponentially and the ideal number of inputs to make the prediction must be found
- Fuzzy C-means clustering which is known for having problems with a high number of inputs.

Chapter 4: Scaling-up of RBF models in bladder cancer prediction

In this chapter, the scaling-up performance of Radial Basis Function Neural-Fuzzy models is investigated. The work presented is based on the challenge of analysing microarray data for the prediction of the patients' cancer survival. The aim is to find: 1) the limit for the maximum number of inputs to use in the model while maintaining low computational complexity and high accuracy. Based on the simulation results presented in this Chapter, the combination of Fuzzy C-means and RBF-Neural-Fuzzy models presents the challenge of scaling-up when more than a thousand inputs are used. To overcome this challenge a Weighted Fuzzy C-means algorithm is introduced. 2) A second contribution is a cluster optimisation algorithm based on the Xie-Beni cluster validity index to improve the quality of the clusters calculated by the WFCM.

4.1 Introduction

The analysis of high dimension-low sample size data represents a systems engineering classification and identification challenge. This is due to the noisy characteristics of high dimensional data and the fact that the number of replications for the experiment is very low (not enough samples for the model's training algorithm to use). The study presented in this chapter is based on the healthcare informatics challenge of analysing large-scale microarray cancer data (high dimension-low sample size data) for the prediction of the patient's cancer survival outcome.

To tackle the challenge of high number of features, feature selection algorithms have become indispensable components of the data mining process [15]. As mentioned in Chapter 2, there are three categories for feature selection: filters, wrappers and embedded methods. Generally, filter feature selection methods are used in combination with wrapper methods to diminish the computational cost of examining the complete data set. The question raised is if the combination of filter and wrapper methods offers significant advantages in terms of tolerance to imprecision and accuracy in the prediction, compared to using only a wrapper method and a higher number of inputs. The combination of filter-wrapper methods have proven to be an effective method for classification [148] .

A number of challenges associated with the theme of this chapter can be addressed; it is important to know if it is possible to avoid the use of feature selection techniques. Specifically, avoid the use of univariate filter-based feature selection techniques that do not assess if there is interdependency in the data, but only assess one-to-one variable dependence. Existing studies suggest that best classification results are obtained by selecting 100-500 genes in a Support Vector Machine model [149, 150]. However, is this limitation a result of the modelling characteristics of SVM models or would a different method provide a better outcome?

In this chapter, an assessment is performed of the scalability of Radial Basis Function Neural-fuzzy models with high dimensionality and low number of samples. An RBF-Neural-Fuzzy system was chosen because it offers a good balance of performance and simplicity while being tolerant to some imprecision and crucially being capable of accurate model representations even when few samples are available [87]. The aim is to assess if it is possible to avoid the use of filter-based feature

selection methods; and conclude if the proposed modelling approach scales-up (i.e. performs well when the number of genes is increased).

As stated in [151]: "One should not rely on clustering results alone for high dimensional data and one should do feature selection". Clustering is a form of data analysis where the data is divided into groups or subsets where the objects present in that subset share some similarities.

Clustering can be divided into two types: hard clustering and fuzzy clustering. Hard clustering refers to an inflexible boundary for the partitions compared to the vagueness showed in fuzzy clustering where a data point may belong to different classes with different membership values [152]. Numerous methodologies have been applied to the problem of clustering HDLSS data, for example: based on p values [153], k-means clustering [154]. In [153] the authors propose a hard clustering algorithm based on p-values as a measure of similarity where no optimisation is necessary. Nevertheless, it is believed that fuzzy clustering is a more appropriate method to find clusters due to its robustness to noise, which is evident in microarray data [120].

In this chapter, the change in the variation of the predictive accuracy of the models, when the number of inputs is increased or reduced, is evaluated using a model for the prediction of survival in bladder cancer [106]. As a pre-input selection the t-test statistical method was used to systematically reduce the large initial dataset. This is a widely applied pre-processing step in microarray gene selection, aiming at eliminating the 'easy to identify' and obviously irrelevant to the process genes.

The method used in the proposed approach, is based on Fuzzy Logic and a Radial-Basis-Function Neural-Fuzzy computational structure. An hybrid Neural-Fuzzy model was chosen for the reason that they have the learning ability of Neural Networks

and the interpretability of Fuzzy Systems. Three different approaches are used in this study to assess the effectiveness of modelling HDLSS data:

a) Fuzzy C-Means and RBF-NF modelling structure;
b) Weighted Fuzzy C-Means and an RBF-NF modelling structure
c) Weighted Fuzzy C-Means and an RBF-NF modelling structure with the help of a cluster validation index.

All the proposed approaches use the Levenberg-Marquardt [144] algorithm for the model's parametric optimisation.

The remainder of this chapter is organised as follows: Section 4.2 Methodology: A description of the data-mining and modelling methodology is presented. Section 4.3 Scaling-up performance of RBF-NF models: Results are shown for the three different modelling approaches applied to the prediction of survival in a bladder cancer, Section 4.4 Analysis of predictive performance and Section 4.5: Summary.

4.2 Methodology

The methodology is organised in three, incremental, parts, whereby a FCM-based RBF-NF modelling approach is presented, then enhanced with measures of weighted-clustering followed by a cluster validity approach.

4.2.1 FCM and RBF-NF function model

The data-mining workflow consists of an initial data pre-processing step, where data normalisation is performed followed by a student's distribution t-test to eliminate easy to identify irrelevant to the process genes. The following step consists of applying Fuzzy C-means clustering for the creation of the initial rule-base. This rule-base is then 'translated' into a Radial-Basis-Function Neural-Fuzzy structure (one multi-dimensional cluster corresponds to one Fuzzy Logic rule), and the modelling structure is finally

parametrically optimised via the Levenberg-Marquardt function-minimisation algorithm [144].

In the same way as in the preceding chapters, the data to be analysed is composed of all the patients and genes plus the survival outcome. Another characteristic of the weighted Fuzzy C-Means is that the number of clusters (rules) needs to be specified, to ensure the low computational complexity; the number of clusters is fixed to 5 (rules). Based on previous research work presented in Chapter 3, five rules in this case study offers a good balance of performance and model simplicity.

4.2.2 WFCM and RBF-NF function model

FCM algorithms consider each object equally important in the cluster solution. For that reason, when FCM is applied to a high number of inputs (more than a thousand), the rule-base loses clarity due to the high dimensional space and the values of the membership degree become truly small. The challenge that arises is that the FCM clusters are the initial conditions for the RBF Neural-Fuzzy and because of their poor quality, the optimisation algorithm fails. By applying Weighted FCM the relative importance of each object to the clustering solution is defined. This weighted factor is applied to the output of the data to improve the membership degree of each cluster. This modification improves the quality of the initial Membership functions of the RBF Neural-Fuzzy model. The second contribution presented in this Chapter (Figure 4.1) consists of applying a Weighted Fuzzy C-means clustering algorithm for the creation of the initial rule-base and applying the rule-base directly to the RBF Neural-Fuzzy model. The rule-base is then ‘translated’ into a Radial-Basis-Function Neural-Fuzzy structure, and is parametrically optimised via the Levenberg-Marquardt function minimisation algorithm [144].

The weighted FCM (WFCM) is based on the minimisation of the following objective function:

$$J_{m=} \sum_{i=1}^{N} \sum_{j=1}^{C} u_{ij}^{m} \; w_i \; \left\| x_i \; - \; c_j \right\|^2, 1 \leq m < \infty \qquad \textbf{(4.1)}$$

where m is any real number greater than 1, u_{ij} is the membership degree of x_i in the cluster j, x_i is the measured data, c_j is the centre of the cluster, and w_i is a weighted factor applied to the output of the data and is equal to the number of inputs.

The membership u_{ij} and the cluster centres c_j are calculated by:

$$u_{ij} \; = \frac{1}{\sum_{k=1}^{c} \left(\frac{\left\| x_i \; - \; c_j \right\|}{\left\| x_i \; - \; c_k \right\|} \right)^{\frac{2}{m-1}}} \; , \qquad \textbf{(4.2)}$$

$$c_j = \frac{\sum_{i=1}^{N} w_i \, u_{ij}^{m} * x^j}{\sum_{i=1}^{N} w_i \, u_{ij}^{m}}$$

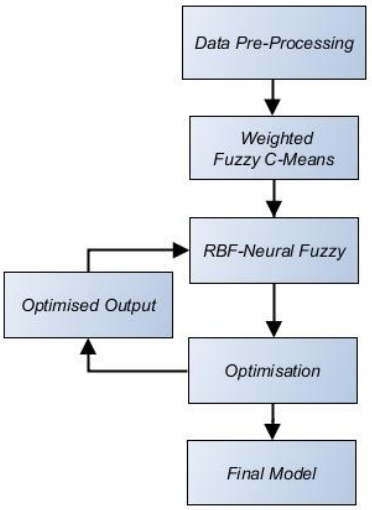

Figure 4.1: Data-mining workflow for the WFCM and RBF-NF model

Each sample will have a membership assigned (u_{ij}) in every cluster; a higher membership would translate into a higher degree of similarity between the sample and the cluster. Each derived information granule (data-cluster) depicts a process rule in the Fuzzy Logic domain. The weighted FCM is similar to the one proposed in [155, 156], however, the novelty of the present work is that the weighting factor changes in relation to the number of genes that are used by the model.

4.2.3 WFCM, validation index and RBF-NF function model

In this section, a cluster-validity index is introduced to the data-mining process to further improve the quality of the rule-base. Figure 4.2 depicts the validity index data-mining workflow. There are multiple indices for validation of the fuzzy clusters; partition coefficient [157], partition entropy[158] , Fukuyama and Sugeno [159], Xie-Beni[160] . Most of the validation indices aim to find the optimal number of clusters, but in this Chapter a modification of the Xie-Beni index is used, as presented in [155], to improve the quality of the clusters calculated by the WFCM. A reliable validation index should take into consideration the compactness or how close each point of the cluster is and the separation of the FCM clusters, which is the case in the Xie-Beni index;

$$Id = \frac{\sum_{k=1}^{N}\sum_{j=1}^{c} w_k (u_{kj})^m \left\| x_k - c_j \right\|^2}{n \, min_{j \neq i} \left\{ \left\| c_j - c_i \right\|^2 \right\}} \quad \textbf{(4.3)}$$

The measure of Compactness (Ct) is given by:

$$Ct = \frac{\sum_{k=1}^{N}\sum_{j=1}^{c} w_k (u_{kj})^m \left\| x_k - c_j \right\|^2}{n} \quad \textbf{(4.4)}$$

The measure of separation is given by:

$$Separation = min_{j \neq i} \left\{ \left\| c_j - c_i \right\|^2 \right\} \quad \textbf{(4.5)}$$

where C is the number of clusters, u_{kj} is the membership degree, w_k is the weight of significance assigned to x_k, which is the complete data, and c_i are the centres of the clusters. The optimal partition clusters would have to be as compact as possible, while they maintain a good balance between separation and coverage of the input space [152]; these characteristics would translate into a high quality rule-base.

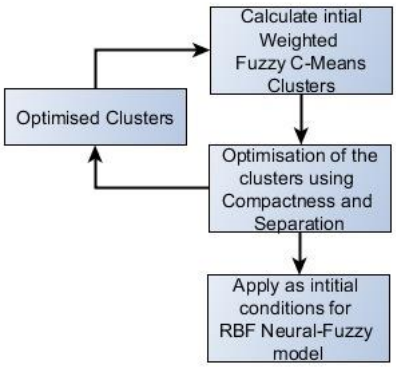

Figure 4.2: Flow chart of the processing of the data with weighted FCM and the validation index

4.3 Scaling-up performance of RBF-NF models

In this section the healthcare-based case study is first introduced, and then the scaling-up simulation results of the modelling methodology are presented.

The introduction of microarray-based technologies for analysing patient tissues has produced a significant challenge for healthcare clinicians as well as data analysis; the challenge of understanding and using efficiently thousands of gene-based data and linking them to clinically useful information. The case study presented in this chapter is focused on bladder cancer microarray data, specifically the ones presented in the Sanchez-Carbayo study [106].

For modelling purposes the survival outcome of the patients was numerically encoded as '-1' for 'No Evidence of Disease – NED' and '1' for 'Dead of Disease - DOD'.

The data samples were randomly separated into 'training' (70% of the patients) and 'testing' (30% of the patients) datasets. The training set is only used to train the model, and the testing dataset is only used after the model training is finished to test the generalisation performance of the model, as a form of cross-validation [147].

In modelling such a dataset, gradually increasing the number of inputs (genes used in the model) also would increase the computational requirements of the process – this may or may not be an issue depending on the application. However, does a larger more complex model (in terms of number of inputs and structure) correspond to an enhanced performance? In the following section (scaling-up performance of RBF-NF models) a comparison between models of 25 to 5000 genes is presented. During this comparison, a number of computational and model performance-related challenges were identified, and it is shown how with the introduction of the proposed data mining and modelling framework helps resolve such challenges. The training time of each model depends on the number of samples and inputs. On average, using a standard PC with an Intel ® Core ™ i7 CPU 870 @ 2.93 GHz processor with 8 GB of RAM, it takes minutes to process (train, test) 25 inputs. The computational requirements increase dramatically, as the number of genes is also increased, to more than 24 hours for 1000 genes. For the models that use 2000 to 5000 genes it was necessary to make use of a High Performance Computing (HPC) server with multiple computing cores [161].

The RBF-NF model was developed as described in section II. The methodology was applied to the data set of Sanchez-Carbayo [106] for the prediction of survival in

bladder cancer. The research question raised at the beginning of this Chapter is if the combination of filter and wrapper methods offers significant advantages in terms of tolerance to imprecision and accuracy in the predictions, compared to using only a wrapper method and a higher number of inputs.

The results shown in this section confirm if it is possible to avoid the use of feature selection techniques. In the case that this premise is true, the rational limit for the maximum number of inputs to use in the model needs to be established. In terms of the modelling structure, five (5) fuzzy rules are maintained throughout the modelling study for comparison purposes. Based on previous research work presented in Chapter 3, five rules in this case study offers a good balance of performance and model simplicity.

Results with 25 inputs and 5 rules

The methodology was applied to the data set of Sanchez-Carbayo to predict the patient's survival of bladder cancer. The results shown in Table 4.1 are the AUC of the models.

Table 4.1: Performance of the model using 25 inputs and 5 rules

Genes	FCM		WFCM		WFCM and validation index	
	AUC		AUC		AUC	
	Train	Test	Train	Test	Train	Test
25	0.96	0.80	0.98	0.63	0.98	0.55

The highest performance is obtained with the FCM model. Also, if these results are compared, both models performed better using FCM. The model with the validation for the initial clusters shows an inferior performance compared to the model that did not used the validation index.

The Gene Signature obtained for the prediction of Survival is shown in Table 4.2. Table 4.2 shows the 25 top ranked genes. For example: Gene RNF1RNF113A is associated to prostate cancer; HLA-A is associated to melanoma; WDR18 is associated with breast cancer; AP3D1 is associated to prostate cancer; ID2 is associated to tumours, cancer and colon carcinoma; PTENP1 is associated to lung cancer; TES is associated to prostate, gastric and ovarian cancer; MCRS1 is associated to breast cancer; GRM8 is associated to prostate cancer; NACC2 is associated to gastric cancer; DNAJC12 is associated to breast cancer; TBCC is related to breast cancer.

Table 4.2: Gene Signature for Bladder Cancer Survival in Sanchez-Carbayo Data Set

Rank	Gene Symbol	Gene Title
1		stage
2	CNR1	cannabinoid receptor 1 (brain)
3	RNF113A	ring finger protein 113A
4	ZHX3	zinc finger homeobox 3
5	HLA-A	major histocompatibility complex, class I, A
6	WDR18	WD repeat domain 18
7	TPST1	tyrosylprotein sulfotransferase 1
8	KLF7	Kruppel-like factor 7 (ubiquitous)
9		grade
10	PRX	periaxin
11	AP3D1	adaptor-related protein complex 3, delta 1 subunit
12	ID2	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein
13	C2orf55	Chromosome 2 open reading frame 55
14	PTENP1	phosphatase and tensin homolog pseudogene 1
15	TES	testis derived transcript (3 LIM domains)
16	MCRS1	microspherule protein 1
17	NR1H3	nuclear receptor subfamily 1, group H, member 3
18	GRM8	glutamate receptor, metabotropic 8
19	FTO	fat mass and obesity associated
20	SLAMF8	SLAM family member 8
21	NACC2	NACC family member 2, BEN and BTB (POZ) domain containing
22	DNAJC12	DnaJ (Hsp40) homolog, subfamily C, member 12
23	TBCC	tubulin folding cofactor C
24	KLHL4	kelch-like 4 (Drosophila)
25	TMEM132A	transmembrane protein 132A

Results with 50 inputs and 5 rules

The results shown in Table 4.3 are the AUC of the models. Once more, the model using FCM was the highest between the three models. The FCM model shows a slight increase in the performance. The model with the WFCM and validation for the initial clusters shows a better performance compared to the WFCM model that did not used the validation index. It also shows a notable improvement compared to the results obtained using 25 inputs.

Table 4.3: Performance of the model using 50 inputs and 5 rules

Genes	FCM		WFCM		WFCM and validation index	
	AUC		AUC		AUC	
	Train	Test	Train	Test	Train	Test
50	0.92	0.83	0.98	0.81	0.96	0.82

The Gene Signature obtained for the prediction of Survival is shown in Table 4.4. Table 4.4 shows the 50 top ranked genes. For example: Gene C18orf8 is associated with prostate and colon cancer; GLI1 is associated with pancreatic cancer; SVIL is associated with prostate cancer; SIK1 is associated with breast and colon cancer; NUCB2 is associated with gastric cancer; PRSS3 is associated with lung and pancreatic cancer; AACS is associated with tracheal cancer; COL16A1 is associated with oral cancer; CEACAM5 is associated with colon cancer.

Table 4.4: Gene Signature for Bladder Cancer Survival in Sanchez-Carbayo Data Set

Rank	Gene Symbol	Gene Title
1		stage
2	CNR1	cannabinoid receptor 1 (brain)
3	RNF113A	ring finger protein 113A
4	ZHX3	zinc finger homeobox 3
5	HLA-A	major histocompatibility complex, class I, A
6	WDR18	WD repeat domain 18
7	TPST1	tyrosylprotein sulfotransferase 1
8	KLF7	Kruppel-like factor 7 (ubiquitous)
9		grade
10	PRX	periaxin
11	AP3D1	adaptor-related protein complex 3, delta 1 subunit
12	ID2	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein
13	C2orf55	Chromosome 2 open reading frame 55
14	PTENP1	phosphatase and tensin homolog pseudogene 1
15	TES	testis derived transcript (3 LIM domains)
16	MCRS1	microspherule protein 1
17	NR1H3	nuclear receptor subfamily 1, group H, member 3
18	GRM8	glutamate receptor, metabotropic 8
19	FTO	fat mass and obesity associated
20	SLAMF8	SLAM family member 8
21	NACC2	NACC family member 2, BEN and BTB (POZ) domain containing
22	DNAJC12	DnaJ (Hsp40) homolog, subfamily C, member 12
23	TBCC	tubulin folding cofactor C
24	KLHL4	kelch-like 4 (Drosophila)
25	TMEM132A	transmembrane protein 132A
26	DUSP2	dual specificity phosphatase 2
27	NUP107	nucleoporin 107kDa
28	CDK14	cyclin-dependent kinase 14
29	MEIS3P1	Meis homeobox 3 pseudogene 1
30	GJA1	gap junction protein, alpha 1, 43kDa
31	LHPP	phospholysine phosphohistidine inorganic pyrophosphate phosphatase
32	FAM208B	family with sequence similarity 208, member B
33	FGF14	fibroblast growth factor 14
34	IBTK	inhibitor of Bruton agammaglobulinemia tyrosine kinase
35	C18orf8	chromosome 18 open reading frame 8
36	GLI1	GLI family zinc finger 1
37	CPLX3	complexin 3
38	NECAB3	N-terminal EF-hand calcium binding protein 3
39	NUP210	nucleoporin 210kDa
40	FAM192A	family with sequence similarity 192, member A

41		receptor (TNFRSF)-interacting serine-threonine kinase 1
42	KIAA1462	KIAA1462
43	SVIL	supervillin
44	SIK1	salt-inducible kinase 1
45	NUCB2	nucleobindin 2
46	PRSS3	protease, serine, 3
47	AACS	acetoacetyl-CoA synthetase
48	COL16A1	collagen, type XVI, alpha 1
49	CEACAM5	carcinoembryonic antigen-related cell adhesion molecule 5
50	COQ7	coenzyme Q7 homolog, ubiquinone (yeast)

Results with 100 inputs and 5 rules

The FCM has the highest performance compared to the other two WFCM models. The difference in performance between the two models using WFCM is however not significant.

Table 4.5:Performance of the model using 100 inputs and 5 rules

Genes	**FCM**		**WFCM**		**WFCM and validation index**	
	AUC		**AUC**		**AUC**	
	Train	Test	Train	Test	Train	Test
100	0.98	0.86	0.96	0.80	0.94	0.79

The Gene Signature obtained for the prediction of Survival is shown in Table 4.6. Table 4.6 shows the 100 top ranked genes. For example: Gene NFX1 is associated with gastric cancer; B3GAT1 is associated with carcinoma; SECISBP2L is associated with lung cancer; GMPS is associated with oral cancer; ST3GAL5 is associated with ovarian cancer; ADRBK2 is associated with colorectal cancer; PEMT is associated with gastric and breast cancer; TMSB10 is associated with breast and ovarian cancer; TAF12 is associated with colorectal cancer; APEH is associated with lung cancer; CLDN1 is associated with breast cancer; FOXN2 is associated with colon cancer; RNF5 is associated with breast cancer; MBOAT7 is associated with bladder carcinoma; TYK2 is

associated with prostate carcinoma; SYDE1 is associated with pancreatic cancer; UBA7 is associated with lung cancer; RPS26 is associated with breast and prostate cancer; LDOC1 is associated with pancreatic cancer; KARS is associated with gastric cancer.

Table 4.6: Gene Signature for Bladder Cancer Survival in Sanchez-Carbayo Data Set

Rank	Gene Symbol	Gene Title
1		stage
2	CNR1	cannabinoid receptor 1 (brain)
3	RNF113A	ring finger protein 113A
4	ZHX3	zinc finger homeobox 3
5	HLA-A	major histocompatibility complex, class I, A
6	WDR18	WD repeat domain 18
7	TPST1	tyrosylprotein sulfotransferase 1
8	KLF7	Kruppel-like factor 7 (ubiquitous)
9		grade
10	PRX	periaxin
11	AP3D1	adaptor-related protein complex 3, delta 1 subunit
12	ID2	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein
13	C2orf55	Chromosome 2 open reading frame 55
14	PTENP1	phosphatase and tensin homolog pseudogene 1
15	TES	testis derived transcript (3 LIM domains)
16	MCRS1	microspherule protein 1
17	NR1H3	nuclear receptor subfamily 1, group H, member 3
18	GRM8	glutamate receptor, metabotropic 8
19	FTO	fat mass and obesity associated
20	SLAMF8	SLAM family member 8
21	NACC2	NACC family member 2, BEN and BTB (POZ) domain containing
22	DNAJC12	DnaJ (Hsp40) homolog, subfamily C, member 12
23	TBCC	tubulin folding cofactor C
24	KLHL4	kelch-like 4 (Drosophila)
25	TMEM132A	transmembrane protein 132A
26	DUSP2	dual specificity phosphatase 2
27	NUP107	nucleoporin 107kDa
28	CDK14	cyclin-dependent kinase 14
29	MEIS3P1	Meis homeobox 3 pseudogene 1
30	GJA1	gap junction protein, alpha 1, 43kDa
31	LHPP	phospholysine phosphohistidine inorganic pyrophosphate phosphatase
32	FAM208B	family with sequence similarity 208, member B
33	FGF14	fibroblast growth factor 14
34	IBTK	inhibitor of Bruton agammaglobulinemia tyrosine kinase

35	C18orf8	chromosome 18 open reading frame 8
36	GLI1	GLI family zinc finger 1
37	CPLX3	complexin 3
38	NECAB3	N-terminal EF-hand calcium binding protein 3
39	NUP210	nucleoporin 210kDa
40	FAM192A	family with sequence similarity 192, member A
41		receptor (TNFRSF)-interacting serine-threonine kinase 1
42	KIAA1462	KIAA1462
43	SVIL	supervillin
44	SIK1	salt-inducible kinase 1
45	NUCB2	nucleobindin 2
46	PRSS3	protease, serine, 3
47	AACS	acetoacetyl-CoA synthetase
48	COL16A1	collagen, type XVI, alpha 1
49	CEACAM5	carcinoembryonic antigen-related cell adhesion molecule 5
50	COQ7	coenzyme Q7 homolog, ubiquinone (yeast)
51	PRAMEF10	PRAME family member 10
52	DET1	de-etiolated homolog 1 (Arabidopsis)
53	NXF1	nuclear RNA export factor 1
54	B3GAT1	beta-1,3-glucuronyltransferase 1 (glucuronosyltransferase P)
55	SECISBP2L	SECIS binding protein 2
56	ACTB	actin, beta /// uncharacterized LOC100505829
57	ASPHD1	aspartate beta-hydroxylase domain containing 1
58	GMPS	guanine monphosphate synthetase
59	RGS9	regulator of G-protein signaling 9
60	ST3GAL5	ST3 beta-galactoside alpha-2,3-sialyltransferase 5
61	ADRBK2	adrenergic, beta, receptor kinase 2
62	PEMT	phosphatidylethanolamine N-methyltransferase
63	GPER1	G protein-coupled receptor 1
64	FOCAD	focadhesin
65	TMSB10	thymosin beta 10
66	TAF12	TAF12 RNA polymerase II, TATA box binding protein (TBP)-associated factor, 20kDa
67	NAGPA	N-acetylglucosamine-1-phosphodiester alpha-N-acetylglucosaminidase
68	MS4A1	membrane-spanning 4-domains, subfamily A, member 1
69	ERVH-6	endogenous retrovirus group H, member 6
70	PYROXD1	pyridine nucleotide-disulphide oxidoreductase domain 1
71	TMCC3	transmembrane and coiled-coil domains 3
72	PALLD	palladin, cytoskeletal associated protein
73	APEH	N-acylaminoacyl-peptide hydrolase
74	CD40	CD40 molecule, TNF receptor superfamily member 5
75	CLDN1	claudin 1
76	FOXN2	forkhead box N2
77	RNF5	ring finger protein 5, E3 ubiquitin protein ligase

78	ZBTB48	zinc finger and BTB domain containing 48
79	KCNK13	potassium channel, subfamily K, member 13
80	GPM6B	glycoprotein M6B
81	MBOAT7	membrane bound O-acyltransferase domain containing 7
82	ZNF259P1	zinc finger protein 259 pseudogene 1
83	TNFSF11	tumor necrosis factor (ligand) superfamily, member 11
84	TYK2	tyrosine kinase 2
85	SYDE1	synapse defective 1, Rho GTPase, homolog 1 (C. elegans)
86	VCL	vinculin
87		disabled homolog 1 (Drosophila)
88	IL11	interleukin 11
89	KLHDC8A	kelch domain containing 8A
90	PPP6R3	protein phosphatase 6, regulatory subunit 3
91	AEN	apoptosis enhancing nuclease
92	UBA7	ubiquitin-like modifier activating enzyme 7
93	COL5A1	collagen, type V, alpha 1
94	RPS26	ribosomal protein S26
95	FAM172A	family with sequence similarity 172, member A
96	ELMO3	engulfment and cell motility 3
97	LDOC1	leucine zipper, down-regulated in cancer 1
98	EXOC5	exocyst complex component 5
99	KARS	lysyl-tRNA synthetase
100	MBTPS1	membrane-bound transcription factor peptidase, site 1

Results with 300 inputs and 5 rules

The results shown in Table 4.7 are the AUC of the model with 300 inputs. In the same manner that the models behave with 100 inputs, the highest performance was obtained by the model using FCM and the difference in performance between the two WFCM models is not significant. Moreover, a trend of increase for the AUC can be perceived for all the models.

Table 4.7:Performance of the model using 300 inputs and 5 rules

Genes	**FCM**		**WFCM**		**WFCM and validation index**	
	AUC		**AUC**		**AUC**	
	Train	Test	Train	Test	Train	Test
300	0.96	0.87	0.98	0.81	1.00	0.82

The Gene Signature obtained for the prediction of Survival contained the 300 top ranked genes. Additional forty two (42) of the included genes that are related with bladder cancer are: Gene THRAP3, TUFM, SERPINB3, GLYR1, TOP2B, FUT6, ICAM2, GLTSCR2, MDC1, C1QBP, MIP, SPG7, CALD1, ITPR1, POLE2, SEC14L2, ITGA5, IL1RN, GRN, FBL, ESRRG, PARP4, MAP2K2, CDH17, NID1, RALA, PCDH7, ISL1, BICD2, EPHX2, MUC3A, FLOT2, PTHLH, APOBEC3A, ASH2L, GOLIM4, ACTN1, GALNT2, ATIC, ALPP, UBC, LIMA1.

Results with 500 inputs and 5 rules

The results shown in Table 4.8 include the AUC of the model with 500 inputs.

Table 4.8: Performance of the model using 500 inputs and 5 rules

Genes	FCM		WFCM		WFCM and validation index	
	AUC		AUC		AUC	
	Train	Test	Train	Test	Train	Test
500	0.94	0.73	0.96	0.75	0.98	0.87

The WFCM model with the validation index for the initial clusters outperforms the WFCM model that did not use the validation index and the FCM model. The WFCM model with the validation index keeps the same trend of an increase in the AUC while the AUC for the WFCM and the FCM model start having a decrease in the performance. The Gene Signature obtained for the prediction of Survival the 500 top ranked genes. Additionally, 37 of the included genes that are related with bladder cancer are: Gene ID3, ZFHX4, POLR2E, OXA1L, BNIP2, PHF17, F2RL2, RBP1, GSTA4, MARS, HEXIM1, NMU, GHITM, IGF1R, NRIP1, IL17RA, MAP2K7, CD2AP, GDF15, CTSE, PENK, IGFBP3, BHMT, PSAP, ELK1, PDE1A, CD3EAP, TFF3, CDKN1C, LONP1, HAL, ALDH4A1, MUC16, CLIC4, AKR1A1, BYSL, TRPC1. 119

Results with 1000 inputs and 5 rules

Table 4.9: Performance of the model using 1000 inputs and 5 rules

Genes	FCM		WFCM		WFCM and validation index	
	AUC		AUC		AUC	
	Train	Test	Train	Test	Train	Test
1000	0.96	0.65	0.96	0.69	0.98	0.80

Similar to the results obtained for the model with 500 inputs, the WFCM model with the validation index for the initial clusters clearly outperforms the WFCM model and the FCM model (Table 4.9). The WFCM model with the validation index now had a decrease in the AUC, the same case presents for the WFMC and the FCM. The Gene Signature obtained for the prediction of contains the 1000 top ranked genes.

Results with 2000 inputs and 5 rules

As discussed earlier in the Chapter, the FCM fails to converge as the complexity increase to more than 2000 genes. This is noted as 'N/A' in the Table 4.10.

Table 4.10: Performance of the model using 2000 inputs and 5 rules

Genes	FCM		WFCM		WFCM and validation index	
	AUC		AUC		AUC	
	Train	Test	Train	Test	Train	Test
2000	N/A	N/A	0.75	0.67	0.75	0.73

The WFCM models have similar performance for training however testing performance is higher for the WFCM model with the validation index. The results also show a general trend of decrease in the performance for both models. The Gene Signature obtained for the prediction of Survival contains the 2000 top ranked genes. Additionally, 181 of the included genes that are related to bladder cancer (Appendix A).

Results with 5000 inputs and 5 rules

Table 4.11: Performance of the model using 5000 inputs and 5 rules

Genes	FCM		WFCM		WFCM and validation index	
	AUC		AUC		AUC	
	Train	Test	Train	Test	Train	Test
5000	N/A	N/A	0.57	0.52	0.57	0.52

As shown in Table 4.11, both models have a significant decrease in the AUC, however they perform the same. Overall, up to around 300 genes, a simple FCM clustering technique is adequate to resolve the modelling complexity of RBF modelling structures. As the number of genes increases, but number of samples remains the same, the WFCM and WFCM with the validity index are needed to model the gene microarray data with a good level of performance. Above 500 genes the WFCM with the validity index starts to outperform the WFCM; however the modelling structure appears to reach its limit in terms of resolving complexity above 5000 genes, where there is a dramatic drop in performance. The Gene Signature obtained for the prediction of Survival contains the 5000 top ranked genes. A total of eight hundred and thirty eight (838) of the included genes that are related to bladder cancer are shown in Appendix A.

4.4 Analysis of predictive performance

Figure 4.3 summarises all the results presented in this section. A clear trend of a decrease in the performance when the number of inputs is increased can be seen. The best performances were obtained when using 300 inputs for FCM and 500 for WFCM with the validation index. Which is consisting to the findings of [150] and [149].

Nevertheless, if the model is produced using less than 300 inputs, the performance is considerably higher for the FCM models; even when using 25 inputs.

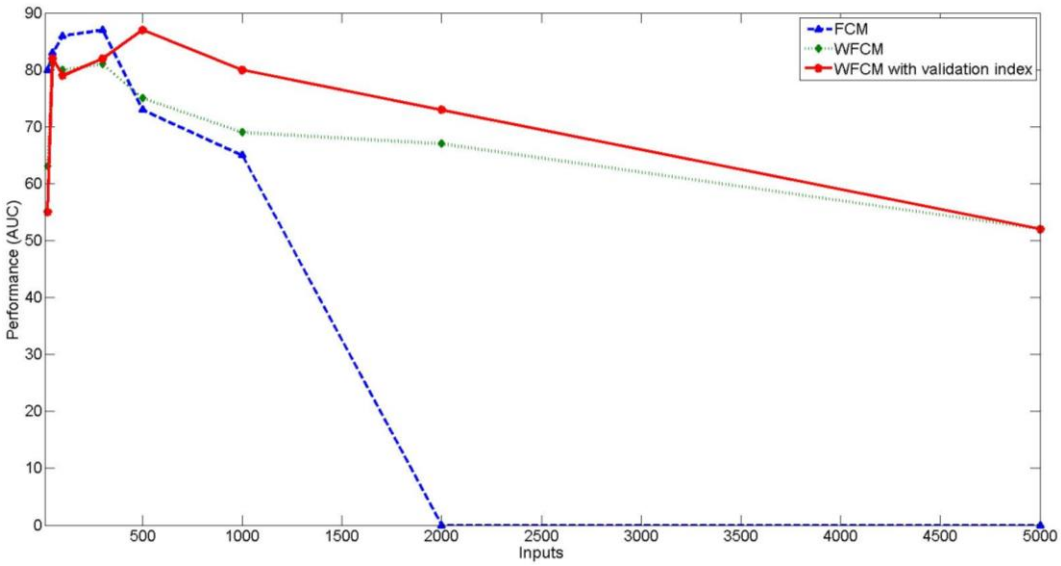

Figure 4.3: Behaviour of the performance for the 3 models.

4.5 Summary

In this chapter, the scaling-up performance of Radial Basis Function (RBF) Neural-Fuzzy models is investigated. RBF-Neural-Fuzzy models offer balance of performance and simplicity (while being tolerant to imprecision); these are traits that are important in healthcare informatics. An enhanced rule-base extraction framework was proposed to improve the model's performance for high-dimensional low sample size data. The work presented in this Chapter is based on the healthcare informatics challenge of analysing large-scale microarray cancer data for the prediction of the patients' cancer survival outcome. The simulations obtained for the prediction of bladder cancer's survival provides a better understanding of the scalability performance for RBF Neural-Fuzzy models.

From the results obtained it can be concluded that the RBF model using FCM alone performs best when less than 300 genes are used. Due to the characteristics of high-dimension low sample size data, as the number of genes increases but number of samples remains the same, the WFCM and WFCM with the validity index are needed to model the microarray data with a good level of performance. Above 500 genes the WFCM with the validity index starts to outperform the WFCM. A dramatic drop in the

performance is observed above 5000 genes, where the modelling structure appears to reach its limit in terms of resolving complexity. Maximum accuracy for the prediction was obtained by using five hundred inputs for the WFCM and the validation index (0.87 AUC) and three hundred inputs for the FCM (0.87 AUC).

The developed models maintain the simple structure with just five (5) rules, but with very good performance (up to 2000 genes). The simple linguistic-based structure of the Fuzzy-logic system could be used in human-centric decision support systems. It is essential to remember that the training time for the models can still be up to 3-4 days on a high performance computing server, however other –more efficient- optimisation algorithm can be used instead.

It must not be forgotten that the models are produced to work closer with clinicians: therefore, apart from a good performance in term of accuracy of AUC the model needs to be comprehensible.

Chapter's summary of achievements:

- Investigate the scaling-up performance of Radial Basis Function Neural-Fuzzy models using a standard PC and a High Performance Computing (HPC) server,
- Find the limit for the maximum number of inputs to use in the model while maintaining low computational complexity and high accuracy.
- An enhanced rule-base extraction framework is proposed to improve the model's performance for high-dimensional low sample size data (microarray data). With the enhanced rule-base, the scaling-up performance of Radial Basis Function (RBF) Neural-Fuzzy models was improved.

The achievements summarised above are linked to one conference publication (International Conference on Computer and Computational Intelligence) and a journal publication in the International Journal of Machine Intelligence and Computing (IJMLC) (post-conference volume-invited).

Based on the modelling structure found on Chapters 3 and 4 (5 rules and 300 inputs using a filter feature selection), a new input selection method will be introduced in the next Chapter; this new method is be based on the polynomial output of the model. The hypothesis behind the New Input selection is to monitor the values of the output weights and membership degree during the training of the structure. Specifically, how the output weights change with every iteration. The assumption is that if the output weight of that particular gene in a certain rule is high that means that it is highly involved in the final output. However, the output not simply is subject to the output weight, the membership degree of certain rule tells us the strength with which that rules is fired. Because of the polynomial output of the model, it is conceivable to distinguish how much a gene is involved in the final output and if that rule is important for the system.

Chapter 5: A new Fuzzy entropy model-based feature selection framework

In this chapter, a new model-based iterative method for feature selection based on fuzzy entropy measures is introduced. The presented approach is based on a Radial Basis Function – Neural Fuzzy modelling structure. A fuzzy entropy measure is used to directly link the relative contribution of each feature to the system's performance. An iterative algorithm is then used for the first time in RBF literature to identify the most relevant features of the process under investigation. In terms of predicting the patients' survival as a result of their bladder cancer gene signature, the inclusion of the cancer stage and grade as extra features of the predictive model is also evaluated, thus producing a hybrid gene-clinical data model. The simulation results confirm that the new approach outperforms existing predictive models in the literature for bladder cancer survival based on gene signature only; the additional novelty of the presented approach relies on the added benefit of producing models that are simpler (considerably less genes in the signature), interpretable, with good generalisation performance and easier to develop and use by non-experts due to the absence of complex pre-processing which is common in this field. The hybrid gene-clinical data model achieves on average 80% accuracy on the prediction of patient survival on "unseen" (new) patient cohorts, confirming the good generalisation of the model. The proposed iterative feature

selection method selects relevant features during the model-training phase, whilst maintaining the system's simplicity and interpretability.

5.1 Introduction

The statement behind the present Chapter is that, as stated in [42]: "a variable that is completely useless by itself can provide a significant performance improvement when taken with others."

As opposed to univariate feature selection, the proposed approach is to generate an embedded model that takes into account the interaction between genes to produce powerful combinations of genes that perhaps are not good by their own, but without overlooking the good prediction performance of the model. In this Chapter, the use of a Radial Basis Function Neural-Fuzzy structure is proposed. The proposed approach consists of an embedded method based on a Radial Basis Function Neural-Fuzzy system [139], which is designed to be equivalent to a Fuzzy Logic Takagi-Sugeno-Kang -based system. A fuzzy entropy measure is used to directly link, for the first time in this modelling structure, the relative contribution of each feature to the system's performance. An iterative model-pruning algorithm is then used to identify the most relevant features of the process under investigation; in this case, a gene signature. The proposed method takes advantage of the link between the output layer of the TSK fuzzy logic modelling structure and each individual rule in the model's rule-base to identify the most relevant genes in each rule of the rule-base. The signature identification is performed in an iterative procedure, thus eliminating the need to pre-process the data-set before developing the process model. The proposed system is tolerant to imprecision, with good generalisation properties and ability to produce accurate predictions even with a low number of features. Because of the low number of features and the simple modelling structure the computational cost is also reduced. The RBF-NF

examines the relationship between gene expression and the outcome (survival, grade or stage) and because is based from fuzzy rules is open for scrutiny and it is possible to understand how the outputs are generated. The biggest strengths of the proposed approach are that the feature selection occurs in the training phase, taking into account the interactions and making it recursive, and that the model is accurate but at the same time interpretable and simple.

The proposed approach is successfully applied to identify suitable gene signatures and predict bladder cancer stage, grade and survival. In three independent data sets [106, 107, 110] the model achieved accuracies ranging from 70% to 99%.

The remainder of this Chapter is organised in four more sections as follows: 5.2 Radial Basis Function for microarray signature identification; 5.3 Entropy Measures; 5.4 RBF Neural-Fuzzy Entropy; 5.5 Simulation Results and 5.6 Summary.

5.2 Radial Basis Function Model for microarray signature

The overall RBF model approach presented in detail in the previous chapter is shown in Figure 3.1, this modelling framework is the core facet of the new Fuzzy Entropy Feature Selection. The new feature selection will be explained in detail in section 5.4 as illustrated in Figure 5.1, the Fuzzy Entropy Feature Selection takes place parallel to the RBF-Modelling and the Levenberg-Marquardt optimisation of the output.

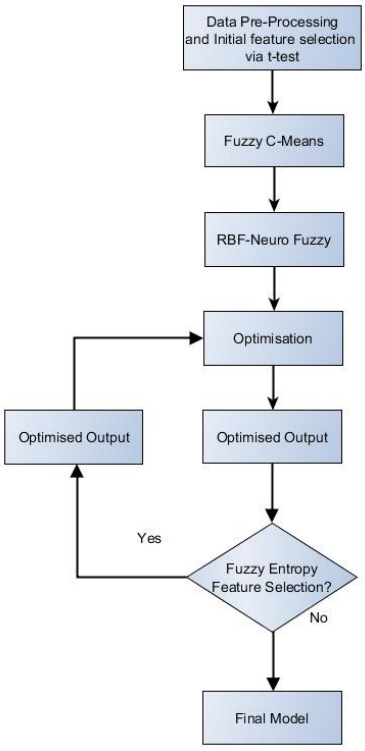

Figure 5.1: RBF-NF Modelling Structure

The datasets are normalised in order to eliminate the high variances between the gene's intensities (quantile normalisation). After normalisation, the student's distribution t-test is used as an initial gene-filter. Based on the p-values the genes from the Sanchez-Carbayo, Blaveri and Kim dataset were reduced from the original number of genes down to a set of 250 genes. Fuzzy C-means algorithm was applied for the elicitation of the initial rule-base. The second stage consists of applying the RBF-NF and optimisation method proposed in Chapter 3.

5.3 Entropy Measures

In this section the concept of entropy as a degree of randomness is used to quantify the fuzziness in a fuzzy system. There is a large dissimilarity between the classical entropy proposed by Shannon that deals with probabilistic uncertainties and the fuzzy entropy that deals with vagueness and ambiguous uncertainties [162].

5.3.1 Definition of Entropy

The introduction of Entropy was made in thermodynamics; it was done by Rudolf Clasius [163] and later expanded by James Clerk Maxwell [164]. The definition of entropy is given by:

$$S=Q/T \tag{5.1}$$

where S is the entropy, Q is the heat content of the system and T is the temperature of the system.

Claude Shannon was one of the first ones to apply entropy outside a thermodynamics or physics. Shannon is acknowledged as the father of Information theory [165]. Information theory deals with the amount of information transferred in an event and is determined by the probability of the event [162], this is referred as quantity of information. It is defined by Equation 5.2:

$$I(A) = -\log P(A) \tag{5.2}$$

where A is an event, $P(A)$ is the probability of the event.

The average of information in all the events is called Entropy [162]. It is typically referred as Shannon's entropy, defined by Equation 5.3:

$$H(X) = -\sum_{i=1}^{n} P_i \log P_i \tag{5.3}$$

where X is a set of variables and P_i is the set of the probabilities in X.

5.3.2 Fuzzy Entropy

Fuzzy entropy is also a measure of information but it is specifically referred as fuzzy information measure [162]. The presented method is based on two features of the RBF-NF model: The Fuzzy Entropy [166] and the Tagaki-Sugeno-Kang (TSK) [125] type of the output layer for the NF system. The fuzzy entropy is calculated via the membership degree of a given input vector to the rule-base of the system. There are various fuzzy entropy measures used in the literature, De Luca and Termini [166] defined the following Fuzzy Entropy measure, which is an average amount of fuzziness and it is based on Shannon's entropy definition; De Luca and Termini introduced a set of properties that Fuzzy Entropy should satisfy:

$$H_{Ai} = -K \sum_{j=1}^{m} \{\mu_j \log(\mu_j) + (1-\mu_j) \log(1-\mu_j)\} \quad \textbf{(5.4)}$$

where:

$H_{Ai} = entropy\ calculated\ per\ rule,\ K = constant$

$m = number\ of\ inputs,\ \mu_j = membership\ degree$

In the current literature, there are examples of use of Fuzzy Entropy in combination with several techniques for feature selection, including: Fuzzy-rough dimensionality reduction [167], microarray and image datasets [123], microarray [70], microarray breast cancer [168], credit scoring [169].

5.4 RBF- Neural-Fuzzy Entropy Feature Selection

As presented in Chapter 3, the RBF modelling structure achieved good results in predicting cancer stage, grade and survival, but one of the best characteristic of this model is the transparency that it give us, translated into high interpretability due to the linguistic rules. In the introduction of this chapter it was discussed how it might be

conceivable to form more powerful combinations by using a combination of genes that do not have a linear or statistic strong dependence to the output, and as an alternative take an approach based on how different combinations of inputs (genes) can give an improved performance. To understand the basic reasoning behind how the New Fuzzy-Entropy feature selection is made it is necessary to monitor the values of the output weights and membership degree during the training of the structure. What is significant to consider is how the output weights change with every iteration, for example if the stay constant or have large variations.

If a weight that relates to a gene stays always constant, that gene may not be significant for the classification stage as it is not "involved" in the training of a particular linguistic rule in the rule-base. On the other hand if one gene fluctuates (either positively or negatively) it may be significant as it contributes towards the prediction strength (entropy) of a particular rule.

The assumption is that if the output weight of that particular gene in a certain rule is high that means that it is highly involved in the final output. However, the output not simply depends on the output weight, the membership degree of certain rule expresses us the strength with which that rules is fired. Because of the characteristics of the RBF-NF model, it is conceivable to identify how strongly a gene is involved in the final output and if that rule is essential for the system.

The method has several challenges to defeat, the number of iterations until a gene is marked as being not significant, or the number of genes the model can handle without losing interpretability and still make a good prediction.

As found in Chapter 4, one of the drawbacks found for FCM clustering is that with high dimensional data, the effectiveness of creating clusters decreases, often

resulting in indistinguishable centres for all of the inputs. The first step is to analyse the dimension that FCM can handle, how many inputs, at what point is not possible to make difference between the clusters. The idea is to monitor the performance of every input of the system and based in the analysis remove the genes that show poor performance.

There are some aspects to take into account:

- The maximum number of inputs used in the model
- Decide if FCM is adequate to calculate the initial clusters
- Optimise the number of rules or clusters
- Best performance of the model , with how many genes
- Selected genes that have medical relevance
- The number of iterations to review the model

The presented methodology is based on two features of the RBF-NF model: The Fuzzy Entropy described in the section 5.3 and the TSK [125] type of the output layer for the NF system described in the previous chapter. The novelty of the approach presented in this chapter is that the fuzzy entropy is used in combination with the antecedent of the RBF fuzzy rule-base (TSK output layer) to assign a relative importance (relevance) to each feature in the dataset (in this case, to each gene). Each single input of the system is analysed in a separate manner to produce a 'factor' that adjusts the aforementioned input based on the behaviour in the system, this is calculated via the fuzzy entropy that takes into account the interaction of the inputs. For the first time in bladder cancer analysis the relevance of input in the model is used to make a ranking of the inputs and a prediction of the cancer's outcome.

While the RBF-NF model optimises the RMSE (see Figure 5.2) via the Levenberg-Marquardt optimisation to make an accurate prediction, the output of the model is analysed to make an input selection at the same time. Providing a distinctive additional characteristic to the embedded model because of the low level of pre-processing needed with the added benefit of maintaining the model's simplicity and low computational cost.

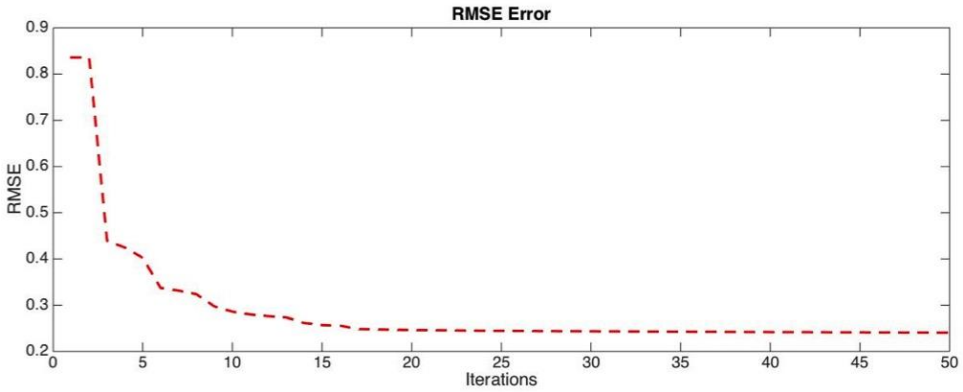

Figure 5.2: RMSE behaviour

It is vital to emphasise that the data was separated into Training (70% of the data) and Testing (30% of the data) to constantly check if the performance parameters show any sign of over-fitting.

The initial step for the Fuzzy Entropy based input selection is to analyse the TSK output layer of an RBF-NF model (see Figure 5.3). The TSK output layer represents a sum of polynomials that are a linear combination of its inputs (Equation 3.8).

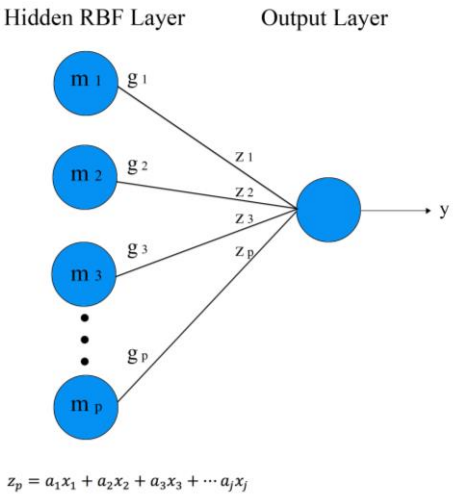

Figure 5.3: TSK output layer of RBF Linear combination of the inputs and each Zi weight is directly linked to g_i rule.

Each TSK weight, z_i, is directly linked to a corresponding rule g_i. The contribution a_j (output weight) of each input x_m can be therefore associated to each rule g_i by the product '$a_j x_j$'. Hence, a relative measure of how important each input is to a specific rule is obtained. Based on the output weights $[a_1, a_2, \dots, a_j]$ it is possible to link how important each feature $[x_1, x_2, \dots, x_j]$ is to a specific rule.

It would be possible to sum all the weights for a specific feature for all the rules to establish a cumulative importance for each feature, however not all the rules of the system contribute in the same way to each prediction. Therefore each rule's g_i(firing strength) and the H_{Ai} (entropy) of each fired rule are used to identify the contribution of each rule to the overall prediction of the model.

The output weights for each rule are then adjusted based on the contribution of each rule, thus the rule-specific significance *'B_{ij}'* of each gene *'j'* per rule 'i' is formulated:

$$B_{ij} = (\bar{g}_i / H_{Ai}) * |a_{ij}| \quad \textbf{(5.5)}$$

where $\bar{g}_i$ is the median value of the firing strength g_i of each rule, and H_{Ai} is the entropy as defined in Eq. (5.4). Therefore, a rule is important if it has a high firing strength (high relevance to the data sample) and low entropy (fuzziness or uncertainty).

The overall importance of each gene (variable) is then calculated by summing up all the rule-specific variable significance (Eq.5.6).

$$B_0 = \sum_{i=1}^{p} B_{ij} \tag{5.6}$$

where B_0 is the overall importance of each gene, B_{ij} is the contribution of each gene in a rule and p is the number of rules. In the algorithmic process proposed here, the model is trained for 'T' iterations, while at 't' iterations (t<T) the training can be 'paused' and the model can be reviewed in terms of the gene ranking order.

Figure 5.4 depicts the overall gene feature selection as a flowchart.

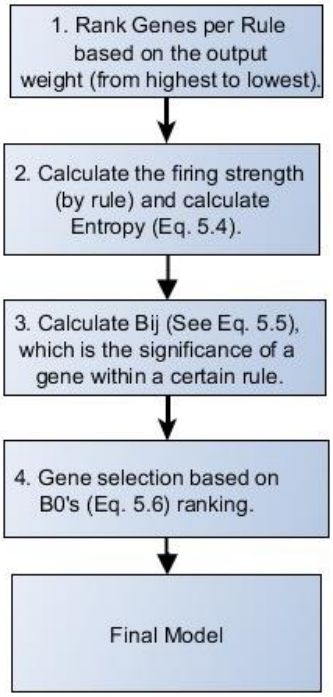

Figure 5.4: Fuzzy Entropy Feature Selection

Described with more detail bellow:

1. The first step is to rank the genes for each rule, based on the output weights (from highest to lowest). Only the top 'n' genes are selected and passed on to the second step. The threshold parameter (to select the top genes) may vary (process-specific). However, as it will be shown in the results section, it was established that, in this study, using twenty five (25) inputs offers a good balance between model simplicity and performance.

2. Calculate the $\bar{g}_i$ and the Fuzzy Entropy H_{Ai} (fuzzy entropy Eq. 5.4).

3. The next step is to calculate the rule-specific significance of each gene B_{ij} (Eq. 5.5). The B_{ij} value represents the behaviour of a gene in a certain rule. For that reason a value of B_{ij} is obtained for each gene in all the rules.

4. Now that the rule-specific significance (B_{ij}value) per gene is obtained, a new ranking is produced. If a gene is shown in several rules its value of B_{ij} is summed up. If a gene is shown in several rules that mean that this specific gene is involved a lot in the final output (hence significant). This measure is not an absolute one; however it provides a relevant measure of significance for the features (genes) in the database, which can be used to provide a feature selection mechanism.

The hypothesis of this algorithmic procedure is similar to using a regression model alone to identify relevant features from the regression coefficients of the polynomial model [170], with the difference here that the RBF-NF consists of multiple polynomials each weighted differently due to its corresponding rule. The advantage of the proposed approach is that here, a model-based approach is considered, where the

importance of each feature is derived by also considering the effect of the rest of the features – as opposed to filter-based approaches where each gene is considered alone, without the effect of the rest of the genes/features.

This procedure is repeated until the desired number of genes is obtained. The model's performance can also be used as a criterion to stop the iterative gene elimination procedure.

During the model training it is possible to observe the cumulative weight of each feature; as the optimisation routine adjusts/optimises the weights of the model the most important genes can be visually identified by the absolute value of their weights (Figure 5.5). The horizontal axis represents the training iterations of the model and the vertical axis the output weights$[a_1, a_2, \dots, a_m]$. Genes, CHPT1, POLE2 and HGFAC are considered important to this example rule as they have a higher contribution compared to Genes, BIC2 or MIP.

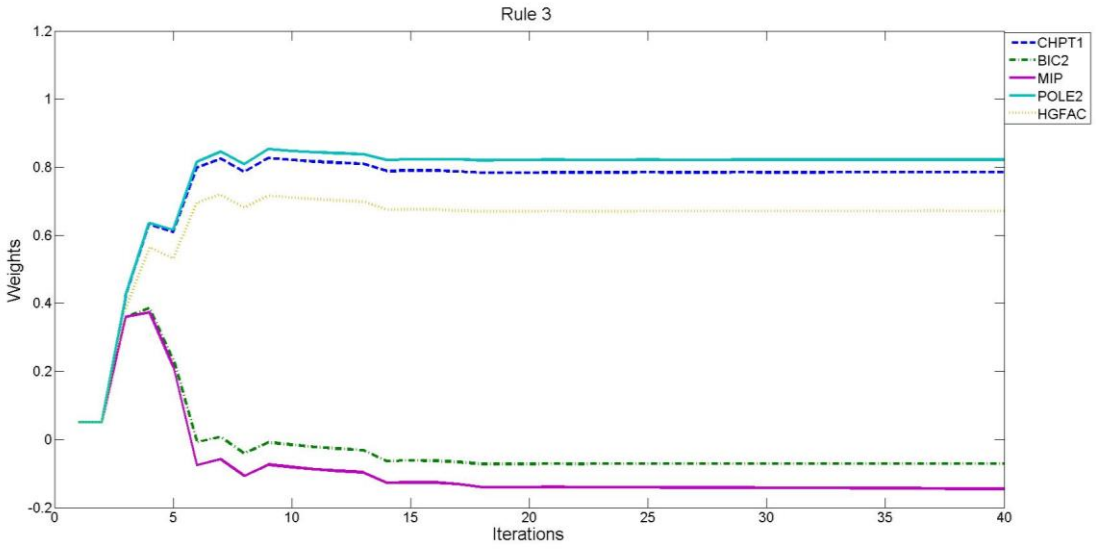

Figure 5.5: Example of the behaviour of the Output weights of 5 Genes in Rule 3.

5.5 Simulation Results

The case-study presented in this Chapter is focused on the prediction of bladder cancer Stage, Grade and patient survival using three different bladder data sets: Sanchez-Carbayo [106], Kim [107] and Blaveri (Table 3.1 Chapter 3). This section is sub-divided into four different parts as follows:

A. Simulation results for Stage and Grade: the model is validated using a real biomedical case-study, which concerns the prediction of the stage and grade of bladder cancer.

B. Simulation results for Survival: a model is produced and validated using the previously mentioned data sets for the prediction of survival of bladder cancer. The present chapter also attempts to identify the best possible combination of clinical data and gene data for the prediction of survival.

C. Fuzzy Logic-type linguistic rule-base: an example of the fuzzy rule-base describing the behaviour of the model.

D. Comparison to existing literature results: the obtained results for the prediction of stage, grade and survival are compared to previously published results.

All the datasets are treated with the same pre-processing procedure as described in the previous Chapter. A pre-selection of the genes was made using the top 250 genes as selected with the t-test. Following the pre-processing stage each dataset is used separately to produce a predictive / feature selection model consisting of 25 inputs, with the results described in the following sections. The training iterations were defined according to the analysis of the results presented in the previous Chapter,

where it was concluded that someplace between 15 to 30 iterations would be sufficient for the model to be trained.

5.5.1 Prediction of patient stage and grade for bladder cancer using microarray data

a) Prediction of patient Stage

The RBF-NF model is developed as described in section 5.2 and 5.4. The methodology was applied to the Sanchez-Carbayo, Kim and Blaveri datasets for the prediction of stage of bladder cancer. The cancer Stage values were 'encoded' into -1 and 1 according to Table 3.4 from Chapter 3. The classification functions of Specificity, Sensitivity and Accuracy are used as measures of performance [147]. The resulting model consisted of 5 rules and 25 inputs. The data samples were randomly separated into 'training' (70% of the patients) and 'testing' (30% of the patients) data-sets. The training set is only used to train the model, and the testing data-set is only used after the model training is finished to test the generalisation performance of the model, as a form of cross-validation [147]. The results shown in Table 5.2 are the mean % of the 10 models for Accuracy, Specificity and Sensitivity respectively. The highest performance was obtained with the Sanchez-Carbayo Data Set, the lowest with Blaveri. Kim had the more balanced performance. For simplicity, only one Gene Signature for the prediction of Stage is shown in Table 5.1. It shows the 25 top ranked genes, the signature has been confirmed from clinicians that is medically relevant.

For example: Gene CEBPD is associated with prostate cancer; ITGB5 is associated with breast and ovarian cancer; HGF is associated with carcinoma; THBS2 is associated with breast cancer and melanoma; RGS1 is associated with melanoma and leukaemia; PVT1 is associated with leukaemia, pancreatic, breast, prostatic and gastric cancer; DUSP1 is associated with ovarian, breast and gastric cancer; SFRP4 is

associated with oral cancer; GADD45B is associated with ovarian cancer and leukaemia; CYR61 is associated with breast and endometrial cancer; NNMT is associated with thyroid, colorectal and gastric cancer; COL10A1 is associated with lung cancer and adenocarcinoma; TAGLN is associated with prostate, colorectal and lung cancer; FLNC is associated with gastric cancer and melanoma; MAPK4 is associated with pancreatic, lung and breast cancer.

Table 5.1: Gene Signature for Bladder Cancer Stage in Sanchez-Carbayo Data Set

Rank	Gene Symbol	Gene Title
1	CEBPD	CCAAT/enhancer binding protein (C/EBP), delta
2	NLGN1	neuroligin 1
3	ITGB5	integrin, beta 5
4	HGF	hepatocyte growth factor (hepapoietin A; scatter factor)
5	THBS2	thrombospondin 2
6	COX7A1	cytochrome c oxidase subunit VIIa polypeptide 1 (muscle)
7	RGS1	regulator of G-protein signaling 1
8	FMO1	flavin containing monooxygenase 1
9	TMEM231	transmembrane protein 231
10	PVT1	Pvt1 oncogene (non-protein coding)
11	DUSP1	dual specificity phosphatase 1
12	SFRP4	secreted frizzled-related protein 4
13	AEBP1	AE binding protein 1
14	CHD8	chromodomain helicase DNA binding protein 8
15	GADD45B	growth arrest and DNA-damage-inducible, beta
16	CYR61	cysteine-rich, angiogenic inducer, 61
17	NNMT	nicotinamide N-methyltransferase
18	RGS2	regulator of G-protein signaling 2, 24kDa
19	COL10A1	collagen, type X, alpha 1
20	TAGLN	Transgelin
21	CYHR1	cysteine/histidine-rich 1
22	TPST1	tyrosylprotein sulfotransferase 1
23	FLNC	filamin C, gamma
24	KANK1	KN motif and ankyrin repeat domains 1
25	MAPK4	mitogen-activated protein kinase kinase kinase kinase 4

In Section 5.5.4 a comparison between the results shown in Table 5.2 and previous publications is presented. The resulting models use a low number of genes (25) and rules (5).

Table 5.2: Prediction of Stage using 5 rules and 25 inputs

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	99	100	98	93	100	89
	Standard Deviation	2	0	3	4	0	7
Blaveri	**Performance (%)**	96	99	91	70	93	67
	Standard Deviation	10	3	27	5	12	5
Kim	**Performance (%)**	90	89	91	76	75	80
	Standard Deviation	3	4	3	6	5	10

b) ***Prediction of patient grade***

The same pre-processing process applied to the data for the prediction of stage was applied. The RBF-NF model was developed as described in Section 5.2 and 5.4. Three grades are used to rate cancer and are encoded according to Table 4.4.

The gene signature for Grade is presented in Table 5.3, only common gene between the two signatures for stage and grade is secreted frizzled-related protein 4, which is associated with oral cancer.

Other genes associated to different types of cancer include: EPHB4 is associated to prostate, ovarian and colon cancer; PCSK5 is associated with colon cancer; STX10 is associated with gastric cancer; AGFG1 is associated with melanoma; SFRP4 as mentioned before is associated with oral cancer; NID2 is associated with ovarian cancer; TMEM184C is associated with thyroid cancer and prostatitis, a disease linked with prostate cancer; CAMK2B is associated with breast cancer; CDC25B is associated with prostate, lung, neck and colon cancer; LAMB4 is associated with lung cancer;

TMPRSS6 is associated with prostatic and breast cancer and prostatitis; NOTCH2 is associated with prostatic and breast cancer and leukaemia; DHRS11 is associated with laryngeal cancer; COL6A3 is associated to colorectal and gastric cancer.

Some of these genes (i.e. SPARC, COL6A3) are related with tumours in general; with the presented model it is conceivable to do an in-depth medical examination if the intensities of these genes are high, acting as an indicator of the malignancy.

Table 5.3: Gene Signature for Bladder Cancer Grade in Sanchez-Carbayo Data Set

Rank	Symbol	Gene Title
1	EPHB4	EPH receptor B4
2	PCSK5	proprotein convertase subtilisin/kexin type 5
3	SLC1A3	solute carrier family 1 (glial high affinity glutamate transporter), member 3
4	STX10	syntaxin 10
5	SYBU	syntabulin (syntaxin-interacting)
6	GPATCH3	G patch domain containing 3
7	AGFG1	ArfGAP with FG repeats 1
8	SFRP4	secreted frizzled-related protein 4
9	NID2	nidogen 2 (osteonidogen)
10	TMEM184C	transmembrane protein 184C
11	RNF141	ring finger protein 141
12	COL11A2	collagen, type XI, alpha 2
13	GPR116	G protein-coupled receptor 116
14	CAMK2B	calcium/calmodulin-dependent protein kinase II beta
15	CDC25B	cell division cycle 25 homolog B (S. pombe)
16	COL5A1	collagen, type V, alpha 1
17	CXCL6	chemokine (C-X-C motif) ligand 6 (granulocyte chemotactic protein 2)
18	LAMB4	laminin, beta 4
19	HIST1H2AJ	histone cluster 1, H2aj
20	TMPRSS6	transmembrane protease, serine 6
21	NOTCH2	notch 2
22	DHRS11	dehydrogenase/reductase (SDR family) member 11
23	SPARC	uncharacterized LOC100505813 /// secreted protein, acidic, cysteine-rich (osteonectin)
24	COL6A3	collagen, type VI, alpha 3
25	FCGR2A	Fc fragment of IgG, low affinity IIa, receptor (CD32)

The results shown in the Table 5.4 are the mean of the 10 models for Accuracy, Specificity and Sensitivity.

Table 5.4: Prediction of Grade using 5 rules and 25 inputs

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	96	98	95	72	77	70
	Standard Deviation	12	6	13	4	8	5
Blaveri	**Performance (%)**	99	95	99	97	93	98
	Standard Deviation	1	16	1	5	5	6
Kim	**Performance (%)**	91	97	84	80	82	77
	Standard Deviation	3	2	7	5	7	12

Contrary to the results obtained for Stage, the performance with the Sanchez-Carbayo data set is the lowest and Blaveri is the highest. Sanchez-Carbayo, Blaveri and Kim had a similar behaviour when the balance in the performance is considered. A more detailed discussion of the results is given in section 5.5.4.

5.5.2 Prediction of patient survival in bladder cancer

This section is focused on the prediction of Cancer Survival; the main focus is to identify the best possible combination for prediction of Survival by combining all the available information from each data set.

In this study, the same RBF Neural-Fuzzy model is applied to the Sanchez-Carbayo, Blaveri and Kim data set to predict the Survival rate. At the same time clinical data (cancer stage and grade) is added to the model to assess if adding such data could enhance the performance or if a prediction can be made using simply clinical data.

a) **Prediction** *of patient survival using only clinical data*

The initial question of this Section is; is it possible to predict survival in bladder cancer using only clinical data? Is this information sufficient to produce an accurate

model? Is it essential the microarray data to generate a prediction? To respond this enquiry a model based only in clinical data must be produced and from the results a hypothesis that might be relevant and aid us enhance the model can be formulated. The model would be a combination of microarray and clinical data. As an initial attempt only Stage was modelled. The results shown in Table 5.5 demonstrate that it is not possible to make an accurate prediction of survival using only this input. If a model has an acceptable performance, the standard deviation is massive. Also the measurements of performance (accuracy, sensitivity and specificity) are not balanced.

Table 5.5: Prediction of Survival using Stage Only

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	72	62	90	62	43	86
	Standard Deviation	8	16	32	2	21	30
Blaveri	**Performance (%)**	58	39	85	36	26	51
	Standard Deviation	2	3	2	4	6	7
Kim	**Performance (%)**	59	78	47	59	80	45
	Standard Deviation	1	34	29	4	31	28

The results shown in Table 5.6 exhibit similar results to the ones found for Stage, it is not possible to make an accurate prediction of Survival using only this input. The performances are either excessively low or less than average but with a large standard deviation.

Table 5.6: Prediction of Survival using Grade Only

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	59	62	60	58	57	58
	Standard Deviation	3	33	52	2	37	50
Blaveri	**Performance (%)**	57	38	84	39	27	55
	Standard Deviation	2	2	1	5	**5**	5
Kim	**Performance (%)**	71	62	89	62	42	86
	Standard	6	13	31	1	20	30

The last attempt is to combine Stage and Grade to make the prediction of Survival. Table 5.7 shows the results for each data set.

It is essential to highlight that regardless of the fact that the performance is not high or better that the previously published results for the same data sets, a certain degree of improvement is shown if the results are compared to Table 5.5 and 5.6. The results show that it is not possible to make an accurate prediction of Survival using only these inputs and that the best model is a combination between Stage and Grade.

Table 5.7: Prediction of Survival using Stage and Grade Only

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	78	68	98	64	42	92
	Standard Deviation	4	8	4	2	4	2
Blaveri	**Performance (%)**	65	92	20	58	86	19
	Standard Deviation	5	15	5	6	27	10
Kim	**Performance (%)**	66	92	20	58	86	40
	Standard Deviation	7	12	42	1	20	19

Based on the results presented in Chapter 3, it can be concluded that the RBF-Neural-Fuzzy model can obtain improved performances than the ones presented in this section, nevertheless, the model can beneficiate from a combination of Stage and/or

Grade to enhance the performance of the model. This assumption will be investigated in Section c) Prediction of Survival in bladder cancer via microarray data and clinical data.

b) Prediction of patient survival via microarray data

An RBF-NF model was developed as described in Section 5.2 and 5.4 using only microarray data intensities. The methodology was applied to the Sanchez-Carbayo, Blaveri and Kim Data-set, to reduce the number of features. The classification functions of Specificity, Sensitivity and Accuracy are used as measures of performance [147]. The developed model consisted of 5 rules and 25 inputs. The data samples were randomly separated into 'training' (70% of the patients) and 'testing' (30 % of the patients) data-sets.

This procedure was repeated ten (10) times as a form of k-fold cross-validation. The results shown in Table 5.8 include the mean of the ten models for Accuracy, Specificity and Sensitivity respectively, along with the standard deviation of each k-fold. The obtained results showed that Sanchez-Carbayo and Blaveri have similar levels of performance, using the exact number of inputs. Kim had a lower and more unbalanced performance. Section b) will try to improve the performances achieved in this section by adding clinical data to the model. Survival was encoded according to Table 3.2 from Chapter 3.

Table 5.8: Survival Prediction using Microarray Data, 5 rules and 25 inputs

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	97	100	94	84	86	79
	Standard Deviation	1	0	3	4	7	8
Blaveri	**Performance (%)**	99	99	99	79	75	83
	Standard Deviation	1	1	1	3	9	5
Kim	**Performance (%)**	86	89	81	67	74	64
	Standard Deviation	2	4	4	6	10	13

c) Prediction of patient survival in bladder cancer via microarray data and clinical data.

Clinical data are then added to the database for the prediction of survival in bladder cancer. The clinical data, which consist of the stage and grade of cancer could improve the modelling performance – providing these new feature are selected by the algorithm.

The cancer Stage and Grade values are ‘encoded’ into -1 and 1 according to Table 3.3 and 3.4.

In comparison, the addition of stage and grade results in improved testing performance (generalisation), which is a very important aspect for survival outcome modelling (the ability to generalise and perform well in unseen data). The three measures of performance, Accuracy, Specificity and Sensitivity (shown in Table 5.10) also appear to be better balanced (evidence of model reliability - robustness). It was also observed that the number of iterations to train the model and the number of features could be reduced while maintaining a very good and balanced overall performance. The resulting gene signature consists of 23 genes, and with the inclusion of the two clinical

features it results to the final 25-input model. The model itself consists of just five (5) rules which represent a very simple modelling structure.

As it shown in the next section, the developed model and signature outperform existing signatures and models in the literature for the same datasets, while the presented signature consists or considerably less genes and/or is much simpler as shown in the comparison in the following section. The identified gene signature is shown in Table 5.9. It has been confirmed with oncology experts that the identified gene signature represents a clinically feasible marker. Some of the selected genes related with cancer are: CHPT1 related with prostatic and breast cancer; CDH16 is related with renal cancer; FGF14 is associated with breast cancer and melanoma; GLI1 is associated with pancreatic and gastric cancer; MDC1 is associated with prostatic, cervical, breast, pancreatic and lung cancer and leukaemia; IGHV5-78 is associated with leukaemia; POLE2 is associated with colorectal cancer; SEC14L2 prostate cancer; HGFAC is associated with prostate, renal, pancreatic cancer; RNF5 is associated with breast cancer; LPHN2 is associated with breast cancer. Similar to the gene signatures for Stage and Grade, there exist several genes that are not certainly linked to any type of cancer but are linked to tumours; this is a remarkable occasion to collaborate with medical expertise and discover new markers related to a type of cancer. A comparable circumstance is presented with a large portion of the proteins included in the gene signature, the model links the prediction of survival to those markers but from a medical perspective the markers are unknown in terms of relation to a certain type of cancer; the same opportunity for the analysis of the markers presents.

Table 5.9: Gene Signature for Bladder Cancer Survival in Sanchez-Carbayo Data Set

Rank	Gene Symbol	Gene Title
1	CHPT1	choline phosphotransferase 1
2	CDH16	cadherin 16, KSP-cadherin
3	PPIAL4A//PPIAL4B//PPIAL4C//PPIAL4G	peptidylprolyl isomerase A (cyclophilin A)-like 4A /// peptidylprolyl isomerase A (cyclophilin A)-like 4B /// peptidylprolyl isomerase A (cyclophilin A)-like 4C /// peptidylprolyl isomerase A (cyclophilin A)-like 4G
4	FGF14	fibroblast growth factor 14
5	GLI1	GLI family zinc finger 1
6	FBL	fibrillarin
7	RGS9	regulator of G-protein signaling 9
8	CACNA1A	calcium channel, voltage-dependent, P/Q type, alpha 1A subunit
9	MDC1	mediator of DNA-damage checkpoint 1
10	CNR1	cannabinoid receptor 1 (brain)
11	KLHDC8A	kelch domain containing 8A
12	ISL1	ISL LIM homeobox 1
13	CALML5	calmodulin-like 5
14	PYROXD1	pyridine nucleotide-disulphide oxidoreductase domain 1
15	IGHV5-78	immunoglobulin heavy variable 5-78 (pseudogene)
16	BICD2	bicaudal D homolog 2 (Drosophila)
17	POLE2	polymerase (DNA directed), epsilon 2, accessory subunit
18	SEC14L2	SEC14-like 2 (S. cerevisiae)
19	KIAA1211L	Chromosome 2 open reading frame 55
20	HGFAC	HGF activator
21	MIP	major intrinsic protein of lens fiber
22	RNF5	ring finger protein 5, E3 ubiquitin protein ligase
23	LPHN2	latrophilin 2
24	stage	
25	grade	

Table 5.10: Prediction of Survival using Stage, Grade and Microarray data, 5 rules and 25 inputs

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	99	98	99	88	89	87
	Standard Deviation	8	1	1	2	3	4
Blaveri	**Performance (%)**	98	100	95	80	73	85
	Standard Deviation	1	0	2	5	11	5
Kim	**Performance (%)**	88	84	91	73	66	79
	Standard Deviation	3	6	4	3	10	8

5.5.3 Fuzzy Logic-type linguistic rule-base

Apart from the very good performance and modelling structure simplicity, the models presented in this chapter maintain a transparent Fuzzy Logic-type linguistic rule-base. Figure 5.6 shows a sample of the rule-base describing the behaviour of the model. For simplicity, just two rules are shown (one for 'negative outcome' and one for 'positive outcome'); these are shown for five out of the 23 genes in the gene signature (complete signature shown in Table 5.9) Two of the linguistic IF-THEN rules that describe the model are shown below to demonstrate the transparency (interpretability) of the modelling method. The corresponding numerical values of the linguistic hedges 'high', 'medium' etc. are determined by the optimisation algorithm via the training data-set. The equivalent linguistic-numerical interpretation of the normalised gene intensity is shown in Table 3.5.

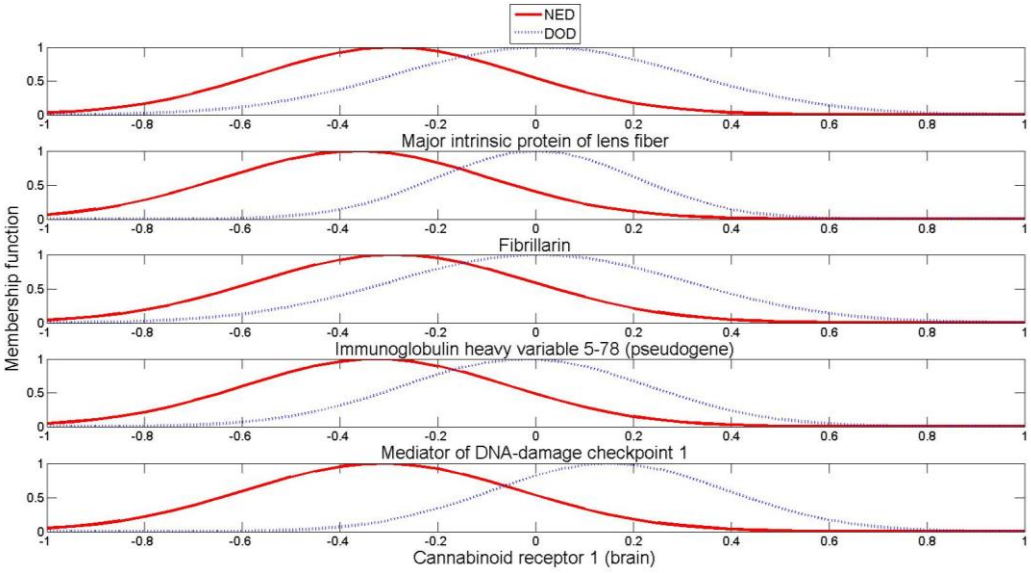

Figure 5.6: Example of a RBF-NF rule base, here for simplicity just two rules are shown, one for a positive outcome and one for a negative.

Rule 2 (NED):

IF the intensity of

the Gene ‘Major Intrinsic protein lens fibre’ is Low Medium and

the Gene ‘Fibrillarin’ is Low Medium and

the Gene ‘Immunoglobulin heavy variable 5-78 finger’ is Low Medium and

the Gene ‘mediator of DNA-damage checkpoint 1’ is Low Medium and

the Gene ‘Cannabinoid receptor 1 (brain)’ is Low Medium

THEN the Patient will survive as results of the disease

Rule 3 (DOD):

IF the intensity of

the Gene ‘Major Intrinsic protein lens fibre’ is Medium and

the Gene ‘Fibrillarin’ is Medium and

the Gene ‘Immunoglobulin heavy variable 5-78 finger’ is Medium and

the Gene ‘mediator of DNA-damage checkpoint 1’ is Medium and

the Gene ‘Cannabinoid receptor 1 (brain)’ is Medium High

THEN the Patient will decease as results of the disease

5.5.4 Comparative Study

In this section the performance of the developed model and associated gene signature are compared to previously published results on the same datasets. Due to the availability of published results and different scope of the various research studies it is not possible to compare all aspects of the presented modelling and feature selection approach.

Therefore, the following comparisons are performed:

- Prediction of stage and grade outcome compared to the Lauss [113] model and the previous model presented in Chapter 3.
- Prediction of survival outcome compared to the Sanchez-Carbayo [106] model and gene signature
- Prediction of survival outcome – muscle invasive tumours only:
    - Compared to the Blaveri [110] model and gene signature
    - Compared to the Riester [114] model and signature.
    - Compared to the model presented in Chapter 3.

a) Comparison of patient stage and grade

Tables 5.8-5.11 show the performance obtained from prediction of Stage and Grade. The RBF Neural-Fuzzy makes accurate predictions but the main advantage is that it is possible to perform input selection at the same time.

The presented model obtained better or comparable performances to previously published results but with the advantage of a ranking of the inputs based in the performance they have in the model [113] with a SVM approach using 150 genes. Table 5.11 shows a comparison between a SVM model with 150 inputs and the RBF model with 25 inputs. The RBF model performed better for the Sanchez-Carbayo data set but for Blaveri Lauss had a better performance. No results for Kim were found to make a comparison.

Table 5.11: Comparison of results from the prediction of Stage to existing publications in the literature

	Lauss (SVM-150 genes) [113]	RBF Neural-Fuzzy (25 genes)
Sanchez-Carbayo	87 %	93 %
Blaveri	85 %	70 %
Kim	-	76 %

Compared to the results shown in Chapter 3 for Stage, the presented model obtained better performances but with the advantage of a ranking of the inputs based in the performance they have in the model and using 25 genes. Table 5.12 shows a comparison between the RBF Neural Fuzzy model with t-test used in Chapter 3 and the RBF Neural Fuzzy model with the Fuzzy Entropy Feature Selection presented in this Chapter. The RBF model performed better or comparably for the Sanchez-Carbayo, Blaveri and Kim data.

Table 5.12: Comparison of results from the prediction of Stage to the results shown in Chapter 3

	RBF Neural Fuzzy with t-test (150 Genes)	RBF Neural-Fuzzy (25 genes)
Sanchez-Carbayo	94 %	93 %
Blaveri	60 %	70 %
Kim	70 %	76 %

Table 5.13 shows a comparison between the Lauss SVM model with 150 genes and the RBF Neural-Fuzzy Entropy model with 25 genes. The RBF model performed better for the Blaveri data set but for Sanchez-Carbayo Lauss had a better performance. No results for Kim were found to make a comparison.

Table 5.13: Comparison of results from the prediction of Grade to existing publications in the literature (Accuracy)

	Lauss (SVM-150 genes) [113]	RBF Neural-Fuzzy (25 genes)
Sanchez-Carbayo	80 %	72 %
Blaveri	86 %	97 %
Kim	-	80 %

Table 5.14 displays a comparison between the RBF Neural Fuzzy model with t-test used in Chapter 3 and the RBF Neural Fuzzy model with the Fuzzy Entropy Feature Selection presented in this Chapter. The RBF model performed comparably for the Blaveri and Kim data. Sanchez-Carbayo had a significant decrease in the performance.

Table 5.14: Comparison of results from the prediction of Grade to the results shown in Chapter 3

	RBF Neural Fuzzy with t-test (150 genes)	RBF Neural-Fuzzy (25 genes)
Sanchez-Carbayo	94 %	72 %
Blaveri	97 %	97 %
Kim	80 %	80 %

b) Survival Outcome Model

In [171], the authors apply Bayesian Networks for predicting the prognosis in breast cancer cases. They showed how the inclusion of the clinical data to the microarray data boosts the modelling performance. In this Chapter, a new model-based

feature selection approach is presented, while showing that the addition of Stage and Grade (clinical data) to gene signature improves performance in the prediction of bladder cancer. The resulting simple structure (five rules and 25 inputs) also aids the computational efficiency of the model.

Table 5.15 shows the performance of the RBF Neural-Fuzzy model compared to existing results from Sanchez-Carbayo [106]. In [25] a SVM modelling structure was utilised with a linear Kernel and the use 250 genes as inputs to the model. The methodology presented in this chapter outperforms the one presented in [25] while, crucially, achieving this with a much simpler structure (25 inputs as opposed to 250). Here only the accuracy measure is compared as sensitivity and specificity measures were not presented in [25].

Table 5.15: Accuracy of Survival using Stage, Grade and microarray data as inputs

	Sanchez-Carbayo (SVM) 250 genes [106]	**RBF Neural-Fuzzy 25 Inputs (23 genes + 2 clinical)**
Sanchez-Carbayo	72 %	88 %

i. Survival Outcome Model - Muscle Invasive Only

The majority of the previously published results for survival in bladder cancer only include muscle invasive cases in order to simplify the modelling approach/structure. Also, from a clinical perspective, these cases are the most important ones to predict in terms of patient survival.

The model developed in this chapter is more generic and includes different stages of bladder cancer. To produce a fair comparison the model was redeveloped with just muscle-invasive patient data. The RBF Neural-Fuzzy model using only muscle invasive patient data was compared to the published results of Blaveri [110], Riester [114] and the results presented in Chapter 3.

Table 5.16 shows the performance of the RBF Neural-Fuzzy model compared to existing results from Blaveri [110] using Prediction Analysis for Microarray (PAM) which uses a modified version of the nearest centroids classification method and 25 genes.

Table 5.16: Performance of Survival (Accuracy) using Stage, Grade and microarray data as inputs

	Blaveri (PAM) [110] (25 genes)			**RBF Neural-Fuzzy (23 genes + 2 clinical)**		
	Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Blaveri	78 %	65 %	93 %	92 %	66 %	100 %

Table 5.17 shows the performance of the RBF Neural-Fuzzy model compared to previous results published by Riester [114]. The Riester study makes use of three independent datasets (Sanchez-Carbayo [106], Blaveri [110]and Kim [107]) to develop a hybrid model using both SVM and a clinical nomogram [115] to assist with the predictions based on 20 inputs. The RBF-NF model exhibits a better balanced performance, Area under the Curve of the Receiver Operating Characteristic (ROC) curve, in two of the three data sets. The RBF-NF model achieves a similar or better performance in all cases with a much simpler modelling structure as the SVM-based model has its predictions are further 'filtered' by a clinical nomogram. The simplicity of the RBF-NF modelling structure might be essential for developing easy to use clinical advisory tools.

Table 5.17: Performance of Survival using Stage, Grade and microarray data as inputs. For comparison purposed the results in this example are shown as the area under the curve (AUC) of a ROC plot

	Survival	
	Riester [114] (SVM + Nomogram 20 genes)	**RBF Neural-Fuzzy 25 Inputs (23 genes + 2 clinical)**
Sanchez-Carbayo	0.74	0.84
Blaveri	0.76	0.83
Kim	0.75	0.72

The results shown in Table 5.18 shows the RBF-NF Entropy model exhibits a better balanced performance (AUC of the ROC curve) in two of the three cohorts. It is important to note that the RBF-NF Entropy model achieves a superior or performance in the Sanchez-Carbayo and Kim case. The addition of stage and grade as an input did increase the performance compared to the results shown in Chapter 3 for the same model using t-test as input selection.

Table 5.18: Performance (AUC) of Survival using Stage, Grade and microarray data as inputs

	Survival	
	RBF Neural-Fuzzy with t-test 20 Inputs	**RBF Neural-Fuzzy 25 Inputs (23 genes + 2 clinical)**
Sanchez-Carbayo	0.82	0.84
Blaveri	0.90	0.83
Kim	0.67	0.72

In summary, the simulation results (Tables 5.11-5.18) show that in most of the cases (where the performance is similar) the accuracy is better than the previously published results for both muscle invasive and non-invasive cases.

One of the main advantages of using the proposed approach is that, as the results demonstrate, via using the RBF-NF approach one can obtain improved or comparable performance but – crucially – via using less number of inputs and with low number of rules (reduced model complexity). The latter also results in very fast model computation

times, in the range of a few minutes when the algorithms are run on a standard single personal computer.

5.6 Summary

This Chapter introduces a new feature selection algorithm based on Fuzzy entropy and a RBF Neural-Fuzzy structure that links directly the fuzzy entropy to the relative significance of the features of the model. Because of the characteristics of the RBF-NF TSK output (input weighted polynomial) a new method is proposed to correlate the features that are more significant to the model's prediction. This significance measure is used to rank the inputs (genes) of the model, via an iterative algorithm.

The proposed methodology has successfully been applied to the case study of bladder cancer prediction for the prediction of the patients' stage, grade and survival outcome.

Another characteristic of this study is how different markers help to predict cancer survival and if they could be used alone (without microarray data). The results show that for the RBF Neural-fuzzy model it is not possible to make an accurate prediction using only Stage and/or Grade but the model benefits from the less noisy nature of that clinical data to generate a more robust prediction output and reduce the number of inputs required to make an accurate prediction. Considering that premise, the combination of clinical data as additional inputs to the most commonly used microarray gene intensities was assessed, finding that the addition of stage and grade improves the overall performance (with various levels of improvement). Crucially, the combination of Stage and Grade and the low number of genes resulted from the approach in this work, helps the model to be developed in simpler structures (low number of rules and

genes, thus reducing model complexity), while maintaining comparable or improved performance as compared to models with significantly more genes or more complex structure. The combination of Stage and Grade also helps the model to reduce the training iterations (easier to optimise), helping to reduce the computational cost to just a few seconds on a standard single personal computer. Via the presented approach a performance equal or better than the work reported in Lauss [113] and Riester [114] is achieved, with the added benefit of the feature selection methodology automatically producing simple models consisting of only 5 rules and 25 inputs (without any significant pre-processing of the data other than the standard normalisation procedures - common for microarray data), with an average performance around 80% success rate in the prediction of patient survival.

Also important in the presented feature selection and modelling approach is the maintenance of the transparency and interpretability of the resulting modelling structure.

The major benefit of this approach, apart from its good accuracy, is the transparency provided by the rule base, converting the rules from the model into a graphical output that can be better understood in a visual manner. Such traits can aid the development of easy to understand and use model by non-experts (non-engineers) such as clinicians in order to directly interrogate the resulting model (human-centric system). Even though the presented methodology was produced for the case study of microarray bladder cancer data, this method may also be applied to numerous other diseases, providing relevant input-output data exist.

Chapter's summary of achievements:

- Development of a Radial-Basis-Function Neural-Fuzzy (RBF-NF) Fuzzy-Entropy based Feature Selection algorithm
- For the first time an embedded RBF-NF model was applied to the feature selection and accurate prediction of stage, grade and survival of bladder cancer.
- The model is shown to maintain its good performance using the inputs selected by the new Fuzzy-entropy feature selection, even when using just 25 genes in the gene based signature.

The achievements summarised above are linked to one conference publication (Biostec 2013, The University of Sheffield- INSIGNEO Institute for In-silico Medicine Showcase, Sheffield, UK (2014) and The University of Sheffield Engineering Symposium - USES 2013, Sheffield, UK (2013).

The present chapter presented the power of the RBF NF network to make accurate predictions even with a low number of inputs. However, all this analysis has been carried on by decreasing the initial data set (consistent of several thousands of genes) using first a filter method (t-test) and then a wrapper or the embedded Fuzzy Entropy feature selection presented in this chapter. The biggest challenge though is presented in the generalisation ability of such data-driven models as identified by other research results too. Models that are trained based on a specific patient cohort should be tested against data from other cohorts to establish the developed models' generalisation performance and predictive robustness. In the next Chapter, the generalisation issues of data-driven models based on microarray analysis will be investigated.

Chapter 6: Generalisation properties of microarray-based models

In this Chapter, the generalisation performance of the developed models is investigated. The approach studied in this chapter is to cross-validate distinct microarray data by applying data integration techniques. Three different data integration approaches were analysed: quantile discretisation, median adjust and NN input-output mapping. The latter two approaches are introduced for the first time to a bladder cancer classification algorithm. The results obtained demonstrate that the data integration methods for cross validation of the models helps to significant increase the predictive performance.

6.1 Introduction

Previous microarray studies have addressed the possibility of comparing different studies (or microarray platforms) [172, 173], concluding that the measurements of gene expression cannot be directly compared but instead the prediction or classification results obtain from this studies can [174]. Most of the research in machine learning algorithms has concentrated on the generation of algorithms able to produce viable classifiers with respect of computational time and generalisation abilities [175]. The challenge investigated in this chapter is why a model that can predict with good accuracy in the same cohort shows poor performance when it is tested on a different data set (cross-validated). It is essential to question if, as stated

in [176]: "on several cases drawbacks in the classifier performance could arise not because of machine learning algorithms, but due to characteristic intrinsic of the data". Could these intrinsic characteristics be solved? And if so, what would the data require for this?

In the present chapter, the possibility of creating a general model that can be used with any type of microarray data set and still make a prediction with good accuracy is investigated. For example, let's consider three data sets: Sanchez-Carbayo-Kim and Blaveri. The first problem arises because the top genes selected by the classifier do not exist in the different data sets. If only the common genes are used, the genes would be a much smaller subset of the original cohort; if the initial number of genes in the data set was ten thousand genes by the time the genes are compared and only the common genes are used, only two thousand and three hundred genes would remain. According to Table 3.1 from Chapter 3, each data set comes from a different platform, and as shown in Figure 6.1 there is no common behaviour between the three data sets. The horizontal axis corresponds to 9 common genes found in the three data sets; the vertical axis corresponds to the gene intensity values normalised from -1 to 1.

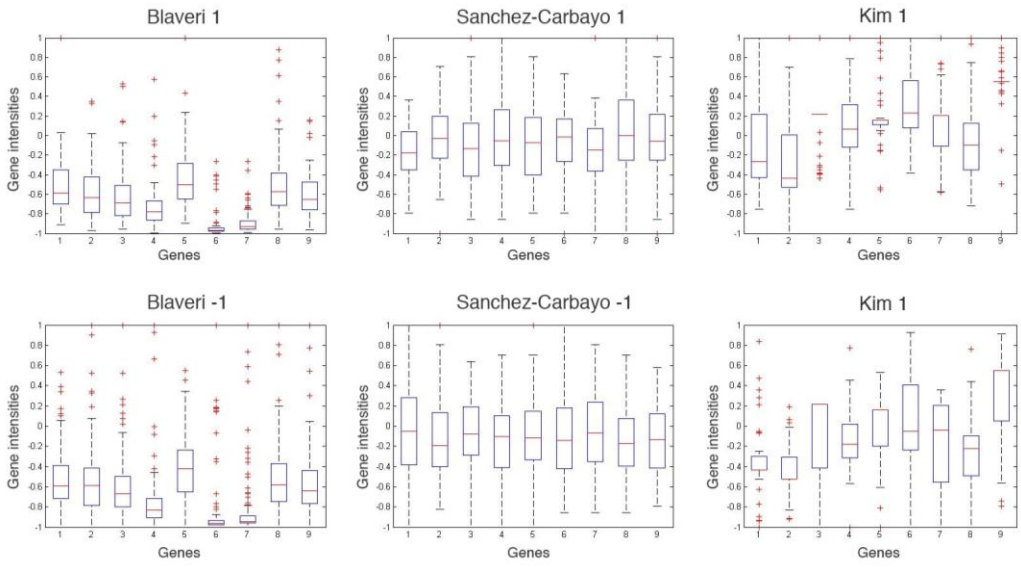

Figure 6.1: Boxplot of behaviour of 3 different data sets. From left to right: Blaveri, Sanchez-Carbayo and Kim.

In Figure 6.1, the intensities are separated into DOD or NED. In Sanchez Carbayo-data set (centre image), a trend of how the median behaves for each of the class can be seen. While for DOD the median value stays closer to -0.2, for NED the median value is closer to 0. This is a clear example of the behaviour expected from a gene; a strong variance in the behaviour from one class to another. In the data set from Blaveri and Kim all the gene intensities behave in a dissimilar manner, there is no clear trend about the behaviour of the data; neither per class or even analysing each gene intensity in the same class.

It is clear that the gene expression intensities need to be processed to be compared directly, this could be due to differences in technologies or in the technique used for the data to be obtained. There are two different approaches to solve this problem: a meta-analysis approach and data integration.

The meta-analysis (Figure 6.2) approach consists of the use of statistical methods to *combine results* from independent studies [177, 178]. The key approach of meta-analysis is to avoid the direct comparison of gene expression values [174].

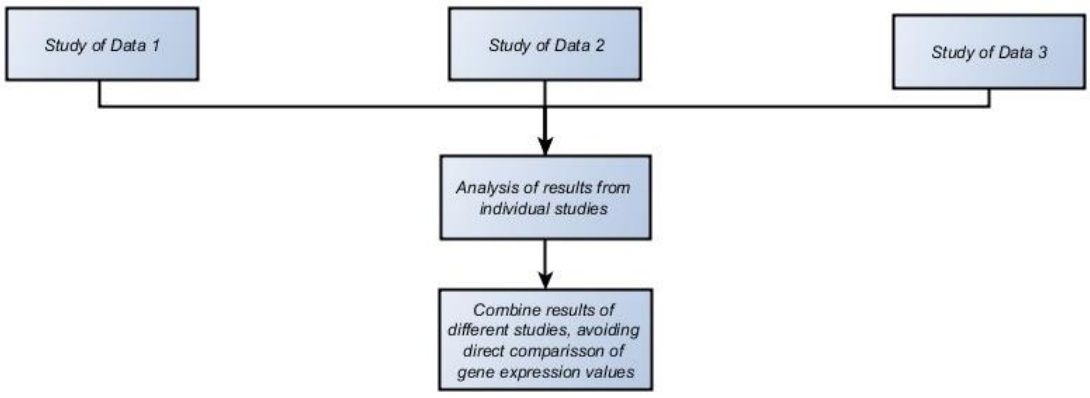

Figure 6.2: Meta-analysis approach

Many different publications have presented meta-analysis results, aiming to rank genes based on confidence measures [179], modelling the unwanted effects of different laboratories [180] or calculating a measure of precision for a study [181]. However, as

stated before, meta-analysis approaches avoid the direct comparison of gene expression values.

An alternative approach is to cross-validate distinct microarray data (Figure 6.3) by applying data integration techniques. Data integration techniques arise due to the high availability of different gene expression data and the opportunity to compare different microarray technologies and cross-validate the results from those experiments. The main challenge is that researchers use different microarray platforms and pre-processing algorithms, making difficult to validate the results found on each data study [174].

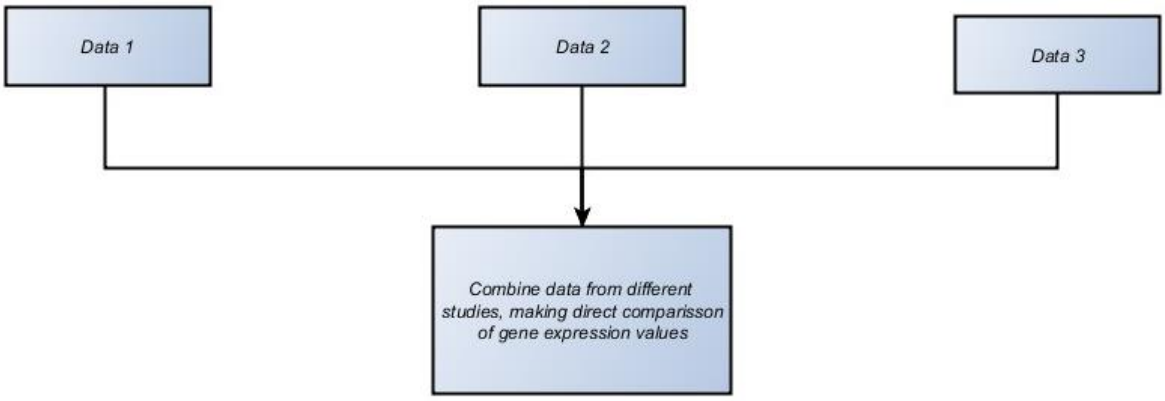

Figure 6.3: Data integration approach

In [182-185] researchers analysed the reproducibility of measurements using different platforms, finding that there is a high reproducibility between the same platforms. There is even a study that introduced a microarray gene expression calibration method [186], however this is only for certain types of microarray chips.

Multiple different data integration methods based on normalisation exist in the literature: [187] proposed a normalisation method using a Z-score, [188] applied rescaling of gene expression values, [189] used normalisation to combine different microarray platforms, [190] proposed a gene scaling factor to integrate microarray data from different platforms.

An alternative approach is to transform the distribution of the data set. This approach was proposed in [174] using quantile discretisation [191] and median rank scores in order to transform the microarray data from different platforms so their distributions become identical. The above mentioned methods already proved to work well for classification tasks but as mentioned in [192], the methods can suffer from information reduction.

It must be emphasised that to date, there is no definitive approach for meta-analysis or data integration because most of the results are data-dependent [177]. One of the biggest challenges is that there is no agreement on which pre-processing algorithm should be used to produce comparable expression measurements across different platforms [193].

The approach investigated in this Chapter, is to cross-validate distinct microarray data by applying three different data integration techniques. The remainder of this Chapter is organised in four more sections as follows: 6.2 Data integration: three different data integration methods are investigated; 6.3 Data Integration Simulation Results and 6.4 Analysis and comparison of results and Section 6.5 Summary.

6.2 Data Integration

The three different data integration methods, presented in this Chapter:

1) median adjust
2) quantile discretisation [191]
3) NN Input-Output mapping.

6.2 .1 Median Adjust

To adjust by the median value of the gene intensities is a common pre-processing step for microarray [113] aiming to centre the gene expression values. In the

approach presented in this Chapter, as an alternative of median centre the data set with its own, a reference data set is used and the median gene expression value for each input are adjust to the median of the reference data set. The median adjust approach is introduced for the first time in this Chapter. This procedure is done by class; in this case two classes (DOD and NED). The resulting data set will have a similar distribution to the reference data; afterwards this process a cross-validation of the models is done.

6.2.2 Quantile discretisation

This method was first applied to microarray breast and prostate cancer [174] and it is based on equal frequency binning [191]. The aim of this method is to discretise the expression values of all arrays into a predetermined number of bins; similar to the analysis investigated in [174], the number of bins is equal to eight for the current investigation. According to the description investigated in [174], for each data set q subsets with equal number of values are determined using the quantiles of the expression value as cut-off points. They defined a cut-off point as the expression value separating an ordered set of expression values into two subsets. The two bins located at the centre are combined into one central bin. The expression values are then substituted by an integer value equivalent to the bin it falls into, a value of zero is given to the central bin and the remaining of the bins are numbered consecutively beginning with the bins next to the centre, using positive integers for the bins containing values above the median and negative values for the rest.

6.2.3 Input-Output Mapping using a Neural Network

The approach presented in this section consists on finding the non-linear relation (mapping) of each input (gene) to its corresponding input from a ‘reference set’. Recapturing the premise expressed at the beginning of the chapter, [176]: “on several cases drawbacks in the classifier performance could arise not because of machine

learning algorithms, but due to characteristic intrinsic of the data". If the analysed data sets are substantially different in terms of gene expression behaviour, but a model that performed with a good accuracy has already been identified, would it be possible to select that data set as a reference and map the gene intensities to its corresponding gene intensity from the 'reference set'? If this premise were correct, the resulting gene intensities would have a prediction performance similar to the one obtained by the 'reference set'. To calculate this relation, a Neural-Network [194] is used. A Neural-Network was used because they are known to be universal approximators, able to approximate any given mapping from inputs to outputs. One of the drawbacks of NN's is that they behave as black boxes but in this case, the extraction of knowledge from the model is not relevant, the objective is to find the input-output relation. To the author's knowledge, no similar approach has been applied for integrating microarray data sets, however an approach presented in [195] applies an ANN (Multi-layer perceptron) to predict functional relationships between proteins.

The NN used in this section is a one hidden-layer feed forward network. An NN consists of several layers; each layer contains a number of units [65]. Figure 6.4 shows the structure of a single hidden layer NN, it consists of an input layer, hidden layer and output layer.

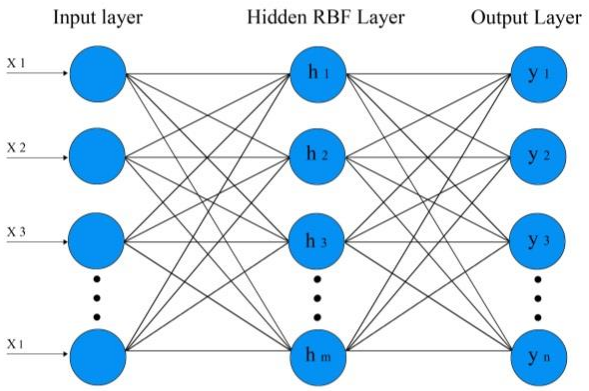

Figure 6.4: One hidden-layer Neural Network

In the example shown in Figure 6.4: l are the inputs, m the hidden layer units and n the outputs. The outputs of the hidden unit are a weighted linear combination of the inputs.

$$a_j = \sum_{i=0}^{l} w_{ji}\, x_i \tag{6.1}$$

where w_{ji} are the weights from the input layer to the hidden layer.

The activation of the hidden layer can be calculated by:

$$h_j = g(a_j) \tag{6.2}$$

The linear combination of the output of the hidden layer is obtained by,

$$a_k = \sum_{j=0}^{m} w_{kj}\, h_j \tag{6.3}$$

Applying the activation function $g2(x)$ to 6.3 the value of the kth output is obtained.

$$y_k = g2(a_k) \tag{6.4}$$

Combining all the equations, the complete representation of the network is:

$$y_k = g2(\textstyle\sum_{j=0}^{m} w_{kj}\ \ g\ (\sum_{i=0}^{l} w_{ji}\, x_i)) \tag{6.5}$$

The methodology applied is simple, to map each input from the reference set to its correspondent input from the validation data sets. Once the corresponding mapping is obtained, the RBF-NF model produced with the reference data set is cross-validated with the corresponding mapped data set of validation data sets. A similar performance to the one obtained by the validation set when only the reference data set was used to produce the model is expected.

As a measure of performance of the model the MSE (Mean square error) is used. The data was separated by classes (NED or DOD) and the inputs were randomly separated into 'training' (70% of the genes), 'testing' (15% of the genes) and 'validation' (15% of the genes) data-sets. A one hidden-layer Neural-Network is used to find the non-linear relation (mapping) of each input (gene) to its corresponding input from a 'reference set'.

6.3 Data Integration Results

In Chapter 5, the results for the prediction of Bladder Cancer's Survival using the fuzzy entropy feature selection method were presented. The first problem arises when validating the gene signature obtained with one data set with the gene signature obtained with a different data set. For example, if the Top 25 Genes obtained from the model using the Sanchez-Carbayo data set are compared with the top 25 Genes obtained with Blaveri's or Kim's data set (Table 6.1 or Appendix B) one would find that none of the genes are present in both lists. This is a generalised problem in microarray analysis because each data set behaves in a different manner [196]. To ensure that the RBF-NF feature selection algorithm was working properly, it was tested it with a benchmark data set, obtaining a median accuracy in selecting the correct inputs of 80% (Appendix C).

Table 6.1: Top Genes for the prediction blader cancer's survival from Sanchez-Carbato, Blaveri and Kim

Rank	Top 25 Gene Title for Sanchez-Carbayo	Top 25 Gene Title for Blaveri	Top 25 Gene Title for Kim
1	choline phosphotransferase 1	hypothetical protein PRO1847	grade
2	cadherin 16, KSP-cadherin	enolase 2, (gamma, neuronal)	stage
3	peptidylprolyl isomerase A (cyclophilin A)-like 4A /// peptidylprolyl isomerase A (cyclophilin A)-like 4B /// peptidylprolyl isomerase A (cyclophilin A)-like 4C /// peptidylprolyl	KIAA0672 gene product	carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 8

	isomerase A (cyclophilin A)-like 4G		
4	fibroblast growth factor 14	transcription factor 15 (basic helix-loop-helix)	adrenomedullin 2
5	GLI family zinc finger 1	zinc finger protein 266	ribosome binding protein 1 homolog 180kDa (dog)
6	fibrillarin	oxytocin receptor	cyclin N-terminal domain containing 2
7	regulator of G-protein signaling 9	tubby like protein 3	lipase, endothelial
8	calcium channel, voltage-dependent, P/Q type, alpha 1A subunit	suppressor of Ty (S.cerevisiae) 4 homolog 1	chromosome 5 open reading frame 46
9	mediator of DNA-damage checkpoint 1	KIAA0410 gene product	espin
10	cannabinoid receptor 1 (brain)	glutamyl-prolyl-tRNA synthetase	phosphodiesterase 6B, cGMP-specific, rod, beta
11	kelch domain containing 8A	syntaxin binding protein 1	transmembrane protein 195
12	ISL LIM homeobox 1	Homo sapiens cDNA FLJ13303 fis, clone OVARC1001372, highly similar to Homo sapiens liprin-alpha4 mRNA	FAT tumor suppressor homolog 1 (Drosophila)
13	calmodulin-like 5	Homo sapiens, clone IMAGE:3940519, mRNA, partial cds	family with sequence similarity 13, member B
14	pyridine nucleotide-disulphide oxidoreductase domain 1	BCL2/adenovirus E1B 19kD-interacting protein 1	N-6 adenine-specific DNA methyltransferase 2 (putative)
15	immunoglobulin heavy variable 5-78 (pseudogene)	Rag D protein	plexin domain containing 2
16	bicaudal D homolog 2 (Drosophila)	Stage	chromosome 1 open reading frame 186
17	polymerase (DNA directed), epsilon 2, accessory subunit	KIAA0027 protein	homeobox and leucine zipper encoding
18	SEC14-like 2 (S. cerevisiae)	proteasome (prosome, macropain) subunit, beta type, 1	chromosome 7 open reading frame 41
19	Chromosome 2 open reading frame 55	guanine nucleotide binding protein 4	aspartylglucosaminidase
20	HGF activator	mitochondrial ribosomal protein L12	similar to programmed cell death 2
21	major intrinsic protein of lens fiber	chromosome 2 open reading frame 8	chloride channel 3
22	ring finger protein 5, E3 ubiquitin protein ligase	Grade	nuclear receptor subfamily 2, group C, member 1
23	latrophilin 2	KIAA0981 protein	N-acetylneuraminate pyruvate lyase 2 (putative)
24	Stage	alanyl-tRNA synthetase	arrestin domain containing 4
25	Grade	Homo sapiens cDNA FLJ10447 fis, clone NT2RP1000851	G protein-coupled receptor 98

From previous results (Chapter 3, 4 and 5) it is possible to conclude that the RBF Neural Fuzzy Model shows a good performance within the same data set. But several experiments must be conducted to conclude that the model can generalise well with a different cohort. To test the generalisation capabilities of the model 3 models were produced (Figure 6.5), the same RBF Neural-fuzzy model is applied to 2300 common genes between Sanchez-Carbayo, Blaveri and Kim data set to predict Survival rate. In addition to the gene intensities, the parameters of cancer Stage classification and cancer Grade classification were considered as inputs to the predictive model. This analysis is investigated in Section 6.3.1.

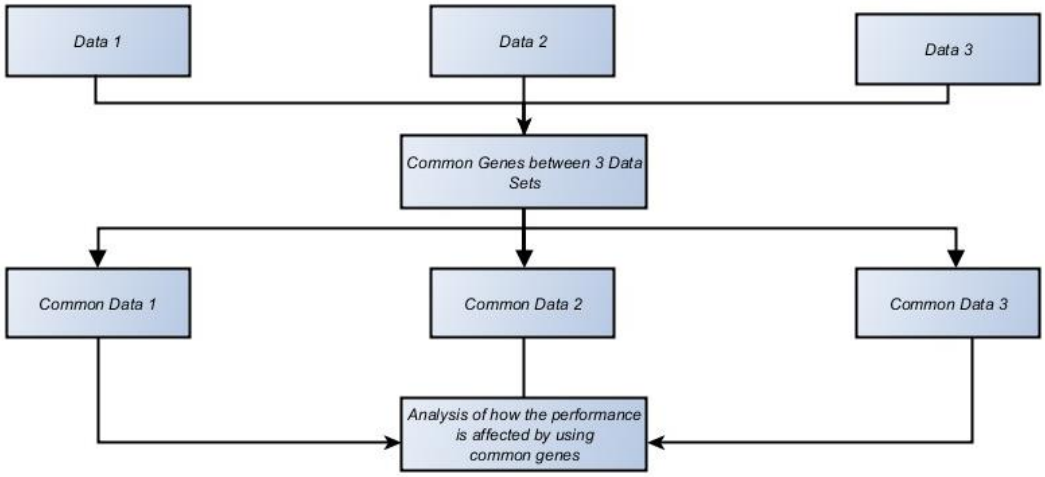

Figure 6.5: Methodology followed for the analysis of the Individual models

The reason to use this approach is to observe how the performance of each individual model is affected by using only the common genes between the three data sets. As shown in Table 6.1, if the three lists of the Top 25 genes selected by the fuzzy entropy feature selection model are compared it is possible to become conscious that none of the genes are repeated in the three data sets. The common genes between the three data sets, which are 2300 in total, are genes that were discarded in the initial stage of pre-processing by t-test. This means that they do not even show a strong linear dependence to the survival outcome. Not only a much smaller subset of the original set remains but also, the quality of the remaining genes is not the ideal to make an accurate prediction. Here, the performance expectation would reduce to around 70% of accuracy.

Once the analysis of how the use of only the common genes to produce a model is done, a reference set will be chosen to cross validate the models and adjusted them to a reference data set. The data set that shows the highest performance across all the results will be chosen as a reference set, which indicates that is the most reliable and with fewer variations in intensities data set.

The remaining 2 data sets will be cross-validated with the reference data set's model and afterwards, three different data integrations methods will be applied (Figure 6.6):

- Median adjusted
- Discretisation
- Mapping Input-Output using a NN

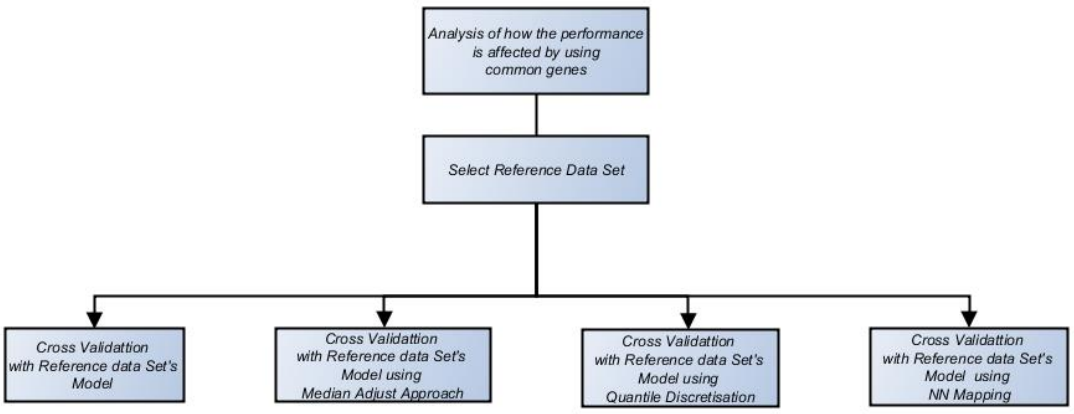

Figure 6.6: Methodology followed for the cross-validation of the models

This analysis is presented in Section 6.3.2

6.3.1 Produce models with common genes

The RBF-NF model is developed as described in section 5.2 and 5.4. The methodology described below was applied to analyse each one of the data sets. The first step was to make a gene input selection using t-test. The gene input selection using t-test was done separately for the 3 data sets and the number of genes was reduced from

2300 genes to 250. The RBF-NF models were trained with the **250 genes** and **5 rules**; afterwards the Fuzzy entropy feature selection algorithm was used on each top 250 genes list. With each data set a model that was trained with the 3 sets of Top 25 Genes selected by the Fuzzy Entropy feature selection was produced. This means that the Sanchez-Carbayo data set was trained 3 times, one with the top genes from his fuzzy entropy feature selection, a second time with the genes selected by Blaveri and a third and final time with the genes from Kim gene selection. The methodology was applied to the Sanchez-Carbayo, Kim and Blaveri datasets for the prediction of survival of bladder cancer. The classification functions of Specificity, Sensitivity and Accuracy are used as measures of performance [147]. The resulting model consisted of 5 rules and 25 inputs. The data samples were randomly separated into 'training' (70% of the patients) and 'testing' (30% of the patients) data-sets. The training set is only used to train the model, and the testing data-set is only used after the model training is finished to test the generalisation performance of the model, as a form of cross-validation [147]. The results shown in Table 6.3 are the mean % of the 10 models for Accuracy, Specificity and Sensitivity respectively. The highest performance was obtained with the Sanchez-Carbayo Data Set, the lowest with Kim. The Gene Signature for the prediction of Survival using the Sanchez-Carbayo data set is shown in Table 6.2. Table 6.2 shows the 25 top ranked genes.

Table 6.2: Gene Signature for Bladder Cancer Survival in Sanchez-Carbayo Data Set

Rank	Gene Symbol	Gene Title
1	FUT6	fucosyltransferase 6 (alpha (1,3) fucosyltransferase)
2	FBL	fibrillarin
3	TOP2B	topoisomerase (DNA) II beta 180kDa
4	CNR1	cannabinoid receptor 1 (brain)
5	MDK	midkine (neurite growth-promoting factor 2)
6	STAT5B	signal transducer and activator of transcription 5B
7	NPTX1	neuronal pentraxin I
8	PTK7	protein tyrosine kinase 7
9	GRIA1	glutamate receptor, ionotropic, AMPA 1
10		grade
11	BAIAP2	brain-specific angiogenesis inhibitor 2
12	PTHLH	parathyroid hormone-like hormone
13	VEGFA	hepatic leukemia factor
14	TNFSF11	tumor necrosis factor (ligand) superfamily, member 11
15	ECE1	endothelin converting enzyme 1
16	GRP	gastrin-releasing peptide
17	TACC2	transforming, acidic coiled-coil containing protein 2
18	TFF3	trefoil factor 3 (intestinal)
19	DGCR2	DiGeorge syndrome critical region gene 2
20	C8A	complement component 8, alpha polypeptide
21	SPAG16	sperm associated antigen 16
22	CEACAM6	carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen)
23	IGFBP3	insulin-like growth factor binding protein 3
24	SH3GL3	SH3-domain GRB2-like 3
25		stage

The results shown in Table 6.3 displays that the RBF-NF Entropy model produced with Sanchez-Carbayo's data set exhibit a superior performance. These results are not surprising since a superior performance for the data set used for making the gene selection is expected, as opposed to the other two data sets that were used for comparison.

Table 6.3: Prediction of Survival using 5 rules and 25 inputs with Sanchez-Carbayo's Top 25 Inputs

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	91	95	86	84	91	71
	Standard Deviation	1	2	4	5	11	21
Blaveri	**Performance (%)**	92	90	94	56	49	60
	Standard Deviation	6	6	7	9	20	19
Kim	**Performance (%)**	87	92	81	54	63	42
	Standard Deviation	3	3	4	4	7	13

Table 6.4 shows the 25 top ranked genes for Blaveri's data set.

Table 6.4: Gene Signature for Bladder Cancer Survival in Blaveri Data Set

Rank	Gene Symbol	Gene Title
1	AARS	alanyl-tRNA synthetase
2	TULP3	tubby like protein 3
3	TCF15	transcription factor 15 (basic helix-loop-helix)
4	CYLC2	cylicin, basic protein of sperm head cytoskeleton 2
5	MYF6	myogenic factor 6 (herculin)
6	DAD1	defender against cell death 1
7	ZNF266	zinc finger protein 266
8	TANK	TRAF family member-associated NFKB activator
9	HAS2	hyaluronan synthase 2
10	SLC4A2	solute carrier family 4, anion exchanger, member 2 (erythrocyte membrane protein band 3-like 1)
11	SNRPB	small nuclear ribonucleoprotein polypeptides B and B1
12	FOXO1	forkhead box D1
13	DNAJB2	DnaJ (Hsp40) homolog, subfamily B, member 9
14	TERF2	telomeric repeat binding factor 2
15	STXBP5	syntaxin binding protein 1
16	ELK1	ELK1, member of ETS oncogene family
17	BTG3	BTG family, member 3
18	NR1H2	nuclear receptor subfamily 1, group H, member 2
19	EPRS	glutamyl-prolyl-tRNA synthetase
20	TPD52L2	tumor protein D52-like 2
21	CUBN	cubilin (intrinsic factor-cobalamin receptor)
22	BCL3	B-cell CLL/lymphoma 3
23	SYN2	synapsin II
24		stage
25		grade

The results shown in Table 6.5 shows the RBF-NF Entropy model produced with Blaveri's data set exhibits a higher Accuracy, if the results are compared to the results obtained using Sanchez-Carbayo's and Kim's data set. Nevertheless, for the results obtained using Blaveri's data set, the Specificity and Sensitivity appear to be unbalanced and with a high standard deviation. The performance obtained by the RBF-NF model using Kim's data set was similar to the one shown in Table 6.3.

Table 6.5: Prediction of Survival using 5 rules and 25 inputs with Blaveri's Top 25 Inputs

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	90	89	91	69	72	64
	Standard Deviation	6	11	3	7	9	21
Blaveri	**Performance (%)**	99	100	98	71	48	80
	Standard Deviation	1	0	2	5	20	9
Kim	**Performance (%)**	86	91	80	54	63	44
	Standard Deviation	6	6	8	6	9	10

The Gene Signature for the prediction of Survival of bladder cancer using Kim's data set is shown in Table 6.6. No common genes between the three data sets (Sanchez-Carbayo, Kim and Blaveri) were found.

Table 6.6: Gene Signature for Bladder Cancer Survival in Kim Data Set

Rank	Gene Symbol	Gene Title
1		grade
2	FGFR4	fibroblast growth factor receptor 4
3	LMNB1	lamin B1
4	SFN	stratifin
5	FOLR3	folate receptor 3 (gamma)
6	ARC	activity-regulated cytoskeleton-associated protein
7	HSD11B2	hydroxysteroid (11-beta) dehydrogenase 2
8	IFI27	interferon, alpha-inducible protein 27
9	DHCR24	24-dehydrocholesterol reductase
10	XRCC3	X-ray repair complementing defective repair in Chinese hamster cells 3
11	TNFRSF9	tumor necrosis factor (ligand) superfamily, member 9
12	CLK3	CDC-like kinase 3
13	TFCP2	transcription factor CP2
14	MAP7	microtubule-associated protein 7
15	CFTR	ATP-binding cassette, sub-family C (CFTR/MRP), member 5
16	CDA	cytidine deaminase
17	PTPN13	protein tyrosine phosphatase, non-receptor type 13 (APO-1/CD95 (Fas)-associated phosphatase)
18	RFX5	regulatory factor X, 5 (influences HLA class II expression)
19	IFIT1	interferon-induced protein with tetratricopeptide repeats 1
20	GABRP	gamma-aminobutyric acid (GABA) A receptor, pi
21	ALDH1A1	aldehyde dehydrogenase 1 family, member A1
22	SNCA	synuclein, alpha (non A4 component of amyloid precursor)
23	CAT	catalase
24	ACVR1	activin A receptor, type I
25		stage

Similar to the results shown in Table 6.5, the results shown in Table 6.7 exhibit a superior performance for RBF-NF model produced using Sanchez-Carbayo's data set. This indicates a trend of superior performance when Sanchez-Carbayo data set is used. No matter which data set was used for making the gene selection, if an RBF-NF model using those genes was produced using Sanchez-Carbayo's data set it would give the

highest validation performance among the three new data sets. These could be due to intrinsic characteristic of the data, the calibration of the instruments at the moment of taking the measurements or simply the processing of the microarray images. This also becomes clear if the distribution of the three data sets shown in Figure 6.1 is analysed. Certainly the same pre-processing of the data was used but they all behave in a different manner. From the results obtained it can be concluded that Sanchez-Carbayo's data set is more 'cleaner' of outlier values and did not have to be filled in for missing values.

Table 6.7: Prediction of Survival using 5 rules and 25 inputs with Kim's Top 25 Inputs

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	96	98	93	75	83	62
	Standard Deviation	2	2	5	6	6	15
Blaveri	**Performance (%)**	98	99	98	55	24	68
	Standard Deviation	2	1	4	13	21	15
Kim	**Performance (%)**	86	92	78	52	63	39
	Standard Deviation	3	5	3	5	11	9

Before cross validating the results into a different data set, a logical approach would be to make a meta-analysis or combination of the results from the Top genes from each data set and produce one 'Top Global Genes List' (shown in Table 6.8). From the results shown in Tables 6.3, 6.5 and 6.7, it is clear that the models produced using Sanchez-Carbayo's data set had the best performance, therefore it would be logic to include a higher number of Top Genes from this data set than from Blaveri or Kim. Since 25 inputs are used in the rest of the models and 2 are already designated for Stage and Grade, it is necessary to divide the rest 23 inputs to give a majority to Sanchez-Carbayo and represent equally Kim and Blaveri. It was decided to use Stage and Grade

plus 13 inputs from Sanchez-Carbayo and 5 inputs from Kim and Blaveri, respectively to give a total of 25 inputs.

Table 6.8: Top Global Genes List

Rank	Gene Symbol	Gene Title
1	FUT6	fucosyltransferase 6 (alpha (1,3) fucosyltransferase)
2	FBL	fibrillarin
3	TOP2B	topoisomerase (DNA) II beta 180kDa
4	CNR1	cannabinoid receptor 1 (brain)
5	MDK	midkine (neurite growth-promoting factor 2)
6	STAT5B	signal transducer and activator of transcription 5B
7	NPTX1	neuronal pentraxin I
8	PTK7	protein tyrosine kinase 7
9	GRIA1	glutamate receptor, ionotropic, AMPA 1
10	BAIAP2	brain-specific angiogenesis inhibitor 2
11	PTHLH	parathyroid hormone-like hormone
12	VEGFA	hepatic leukemia factor
13	TNFSF11	tumor necrosis factor (ligand) superfamily, member 11
14	AARS	alanyl-tRNA synthetase
15	TULP3	tubby like protein 3
16	TCF15	transcription factor 15 (basic helix-loop-helix)
17	CYLC2	cylicin, basic protein of sperm head cytoskeleton 2
18	MYF6	myogenic factor 6 (herculin)
19	FGFR4	fibroblast growth factor receptor 4
20	LMNB1	lamin B1
21	SFN	stratifin
22	FOLR3	folate receptor 3 (gamma)
23	ARC	activity-regulated cytoskeleton-associated protein
24		grade
25		stage

As shown in Table 6.9, there is no significant advantage of using this Top Global Gene list, in fact the performance for Sanchez-Carbayo's model decreased considerably for the Specificity. While Kim and Blaveri showed lower and unbalanced performances.

Table 6.9: Prediction of Survival using 5 rules and 25 inputs with Top Global Gene List

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	92	89	96	70	45	83
	Standard Deviation	16	30	1	3	17	8
Blaveri	**Performance (%)**	86	90	93	47	76	23
	Standard Deviation	1	2	3	3	7	5
Kim	**Performance (%)**	93	92	92	51	70	43
	Standard Deviation	4	7	8	12	13	19

In the next sections the effect of integrating data sets using Sanchez-Carbayo's data set as a reference to try to 'adjust' Blaveri and Kim data sets to improve the Generalisation performance (Testing in unseen and cross validated data set) is investigated.

6.3.2 Cross-validate models

Since the Top Global Genes list did not bring any benefit in performance it was decided to simplify the methodology and use only one data set as the reference. To cross validate the models and adjusted them to a reference data set, the Sanchez-Carbayo's data set was chosen. This data set showed the highest performance across all the results, which indicates that is the most reliable and with fewer variations in intensities data set.

The first approach consists of using the RBF-NF model developed with Sanchez-Carbayo's data set. The data samples were randomly separated into 'training' (70% of the patients) and 'testing' (30% of the patients) data-sets. The training set is only used to train the model, and the testing data-set is only used after the model training is finished to test the generalisation performance of the model, as a form of cross-validation. Instead of using the Testing data from Sanchez-Carbayo, Blaveri and

Kim's complete data are used as Testing to review the generalisation of the model. Additionally, three different data integrations approaches are analysed:

- Median adjusted
- Discretisation
- Input-Output Mapping using a NN

a) Cross-Validate Results

As explained above, the methodology is to cross-validate (use as Testing) Blaveri and Kim's complete data set to review the generalisation of the model. The data from Blaveri and Kim are cross validated with model created with Sanchez-Carbayo's data. It must be emphasised that all the data sets had the same pre-processing. Table 6.10 shows the performance obtained when using Sanchez-Carbayo's data set to produce the model.

Table 6.10: Prediction of Survival using 5 rules and 25 inputs with Sanchez-Carbayo's Top 25 Inputs

		Training			Testing		
		Accuracy	**Specificity**	**Sensitivity**	**Accuracy**	**Specificity**	**Sensitivity**
Sanchez-Carbayo	**Performance (%)**	91	95	86	84	91	71
	Standard Deviation	1	2	4	5	11	21

Table 6.11 and 6.12 shows the Testing performance obtained when using Blaveri and Kim's data set. Both accuracies were considerably low.

Table 6.11: Prediction of Survival using 5 rules and 25 inputs with Sanchez-Carbayo's Top 25 Inputs Cross validated with Blaveri as Testing

		Testing model using Blaveri data set		
		Accuracy	**Specificity**	**Sensitivity**
Blaveri	**Performance (%)**	42	76	14
	Standard Deviation	2	5	7

Table 6.12: Prediction of Survival using 5 rules and 25 inputs with Sanchez-Carbayo's Top 25 Inputs Cross validated with Kim as Testing

		Testing model using Kim data set		
		Accuracy	**Specificity**	**Sensitivity**
Kim	**Performance (%)**	56	93	05
	Standard Deviation	3	17	17

b) Median Adjusted

The median adjust approach is introduced for the first time in this Chapter. As an alternative of median centre the data set with its own median gene expression value for each input (gene), a reference data set (Sanchez-Carbayo) is used and the different data sets are adjust to the median value of the input from reference data set. This procedure is done by class; in this case two classes DOD and NED are used. The resulting data set will have a similar distribution to the reference data, afterwards this process the models are cross-validated per data set, in a similar way as the cross-validation done in the previous sub-section of this chapter.

The performance presented in Table 6.13 is similar to the performance seen when the Blaveri data set was used as Testing without any processing of the data for the model produced with Sanchez-Carbayo's data.

Table 6.13: Prediction of Survival using 5 rules and 25 inputs with Sanchez-Carbayo's Top 25 Inputs with Data Integration Cross-validated with Blaveri

		Testing model using Blaveri data set		
		Accuracy	**Specificity**	**Sensitivity**
Blaveri Median Adjusted	**Performance (%)**	53	55	51
	Standard Deviation	5	8	3

Similar results were obtained when Kim's data set was used for testing the model produced with Sanchez-Carbayo's data. An increase in the performance is seen for the median adjusted model compared to the performance presented in Table 6.14.

Table 6.14: Prediction of Survival using 5 rules and 25 inputs with Sanchez-Carbayo's Top 25 Inputs with Data Integration Cross-validated with Kim

		Testing model using Kim data set		
		Accuracy	**Specificity**	**Sensitivity**
Kim Median Adjusted	**Performance (%)**	62	82	33
	Standard Deviation	2	12	10

By adjusting the median value across all the samples the data sets are forced to have a similar distribution to Sanchez-Carbayo's data set (Figure 6.7, 6.8 and 6.9). The horizontal axis represents the gene intensities and the vertical axis the number of samples. Before any processing of the data, all the data sets had the same pre-processing.

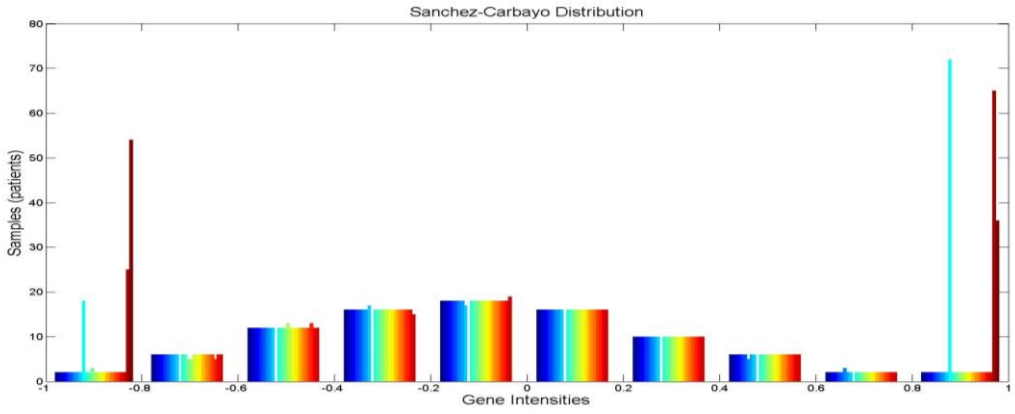

Figure 6.7: Median adjusted for the three data sets using as reference Sanchez-Carbayo. Distribution from Sanchez-Carbayo data set.

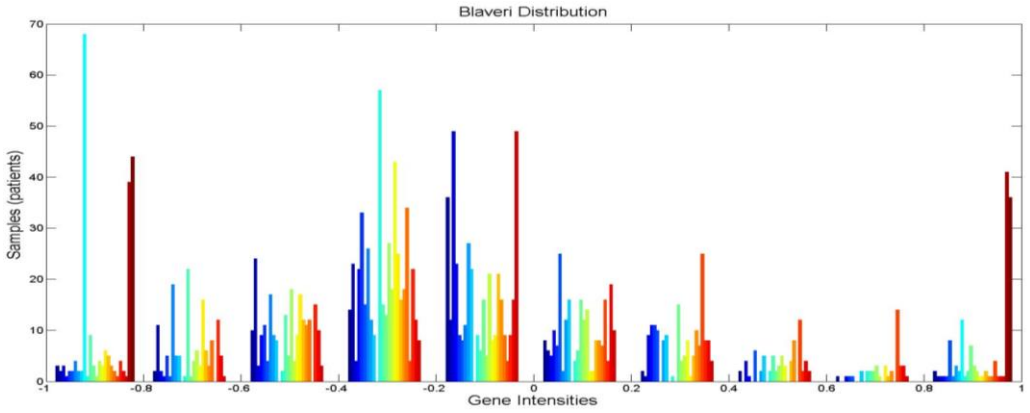

Figure 6.8: Median adjusted for the three data sets using as reference Sanchez-Carbayo. Distribution from Blaveri data set.

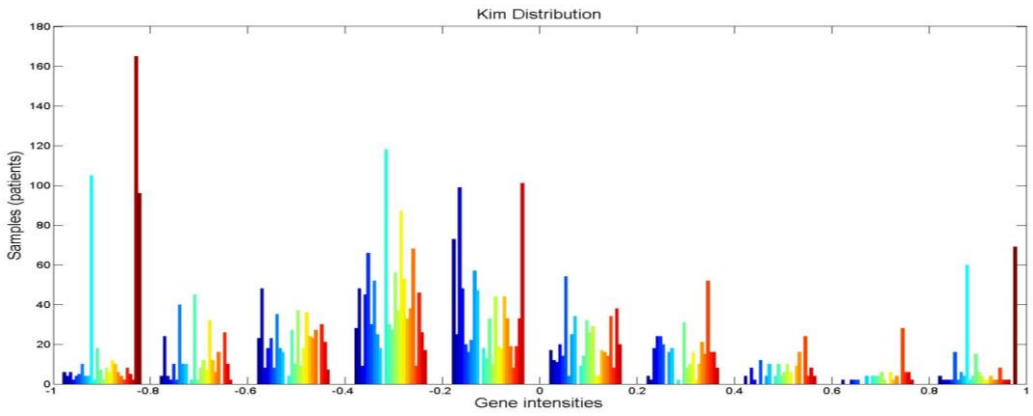

Figure 6.9: Median adjusted for the three data sets using as reference Sanchez-Carbayo. Distribution from Kim data set.

c) Quantile discretisation

The aim of this method is to discretise the expression values of all arrays into a predetermined number of bins. For each data set q subsets with equal number of values are determined using the quantiles of the expression value as cut points. They defined a cut-off point as the expression value separating an ordered set of expression values into two subsets. The two bins located at the centre are combined into one central bin. The expression values are then substituted by an integer value equivalent to the bin it falls into, a value of zero is given to the central bin and the remaining of the bins are numbered consecutively beginning with the bins next to the centre, using positive integers for the bins containing values above the median and negative values for the rest.

As shown in Table 6.15, the performance from Sanchez-Carbayo when the quantile discretisation method is applied was slightly lower compared to the results presented in Table 6.10.

Table 6.15: Prediction of Survival using 5 rules and 25 inputs with Sanchez-Carbayo's Top 25 Inputs with Data Integration

		Training			**Testing**		
		Accuracy	Specificity	Sensitivity	Accuracy	Specificity	Sensitivity
Sanchez-Carbayo Discretisation	Performance (%)	98	97	99	76	77	74
	Standard Deviation	1	1	1	7	8	16

When the Blaveri data set is used as Testing for the model produced with Sanchez-Carbayo's data, a higher performance was achieved using the discretisation method (Table 6.16) as opposed as the results shown for median adjust (Table 6.13).

Table 6.16: Prediction of Survival using 5 rules and 25 inputs with Sanchez-Carbayo's Top 25 Inputs with Data Integration Cross-validated with Blaveri

		Testing model using Blaveri data set		
		Accuracy	**Specificity**	**Sensitivity**
Blaveri Discretisation	**Performance (%)**	61	50	71
	Standard Deviation	7	11	12

Similar results are obtained using Kim data set (Table 6.17); it can be seen an increase in the performance for the quantile discretisation model compared to the performance investigated in Table 6.14.

Table 6.17: Prediction of Survival using 5 rules and 25 inputs with Sanchez-Carbayo's Top 25 Inputs with Data Integration Cross-validated with Kim

		Testing model using Kim data set		
		Accuracy	**Specificity**	**Sensitivity**
Kim Discretisation	**Performance (%)**	62	68	53
	Standard Deviation	2	6	3

By applying quantile discretisation, the data sets are forced to have a similar distribution to Sanchez-Carbayo's data set (Figure 6.10, 6.11 and 6.12). The horizontal axis represents the gene intensities and the vertical axis the number of samples.

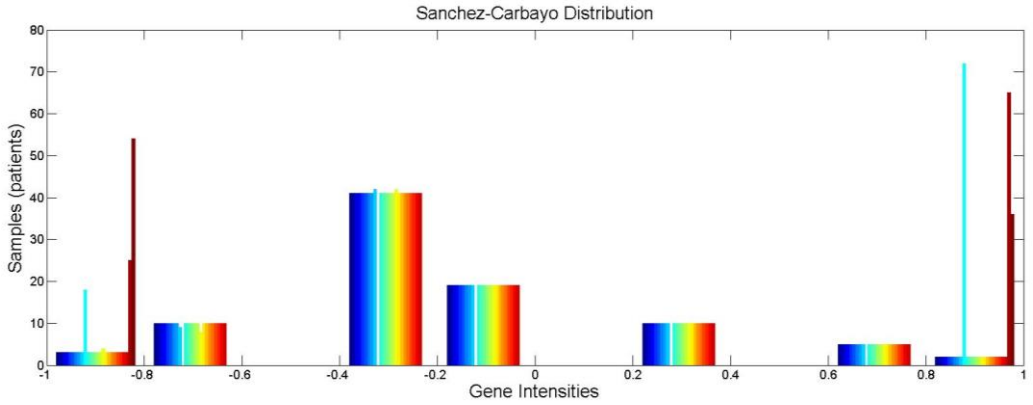

Figure 6.10: Quantile discretisation for the three data sets using as reference Sanchez-Carbayo. Distribution from Sanchez-Carbayo Data set data set.

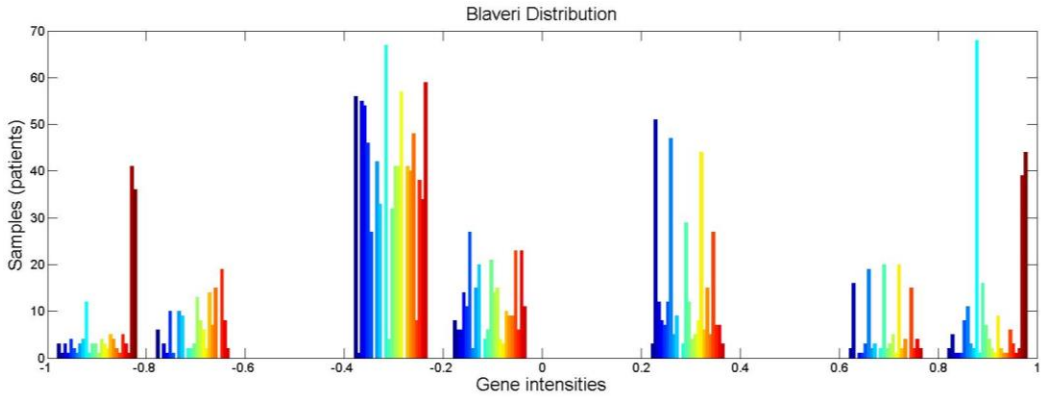

Figure 6.11: Quantile discretisation for the three data sets using as reference Sanchez-Carbayo. Distribution from Blaveri data set.

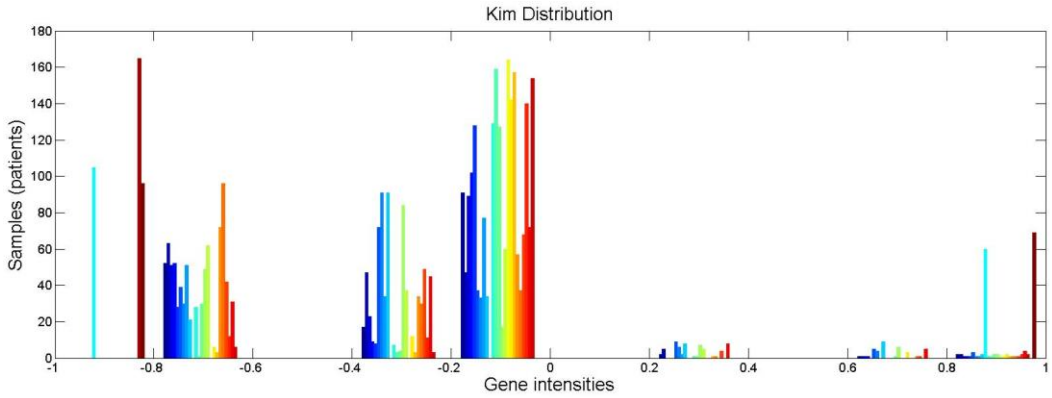

Figure 6.12: Quantile discretisation for the three data sets using as reference Sanchez-Carbayo. Distribution from Kim data set.

d) Input-Output Mapping using a Neural-Network

The Sanchez-Carbayo data set was chosen as the reference data set to map the input-output mapping of two data sets. When the Blaveri data set was used as Testing for the model produced with Sanchez-Carbayo's data a similar performance to the one presented in Table 6.10 is seen. The accuracy, specificity and sensitivity presented in

Table 6.18 are comparable to the corresponding values of Accuracy, Sensitivity and Specificity presented in Table 6.10. The standard deviation is higher for the accuracy performance, which means that there was more variation in performance between the 10 folds.

Table 6.18: Prediction of Survival using 5 rules and 25 inputs with Sanchez-Carbayo's Top 25 Inputs with Data Integration Cross-validated with Blaveri

		Testing model using Blaveri data set		
		Accuracy	**Specificity**	**Sensitivity**
Blaveri	**Performance (%)**	85	94	77
	Standard Deviation	11	16	10

Figures 6.13 and 6.14 display the behaviour of the MSE used as a measure of performance of the model. The best validation performance is obtained at 2 epochs for NED class and 3 epochs for DOD.

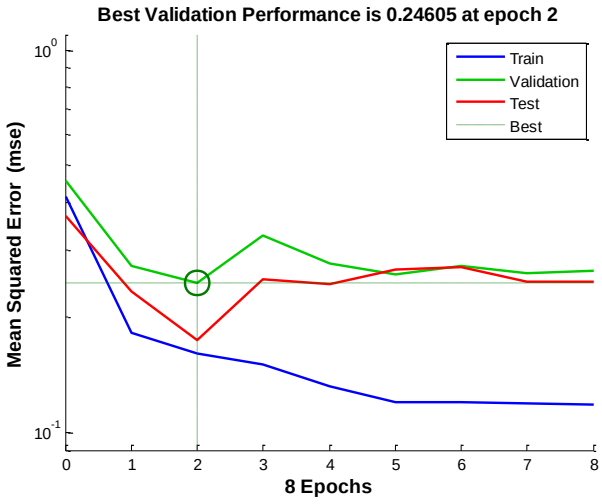

Figure 6.13: Class NED best validation performance

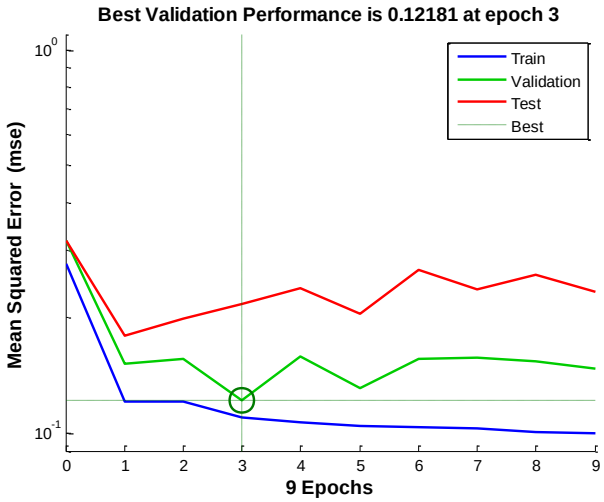

Figure 6.14: Class DOD best validation performance

Similar results are obtained when Kim's data set is used for testing the model produced with Sanchez-Carbayo's data. It is possible to perceive an increase in the performance compared to the results presented in Tables 6.14 (median adjusted) and 6.17 (quantile discretisation) compared to the performance investigated in Table 6.19.

Table 6.19: Prediction of Survival using 5 rules and 25 inputs with Sanchez-Carbayo's Top 25 Inputs with Data Integration Cross-validated with Kim

		Testing model using Kim data set		
		Accuracy	**Specificity**	**Sensitivity**
Kim	**Performance (%)**	79	92	62
	Standard Deviation	17	23	22

Figures 6.15 and 6.16 display the behaviour of the MSE used as a measure of performance of the model. The best validation performance is obtained at 6 epochs for NED class and 5 epochs for DOD.

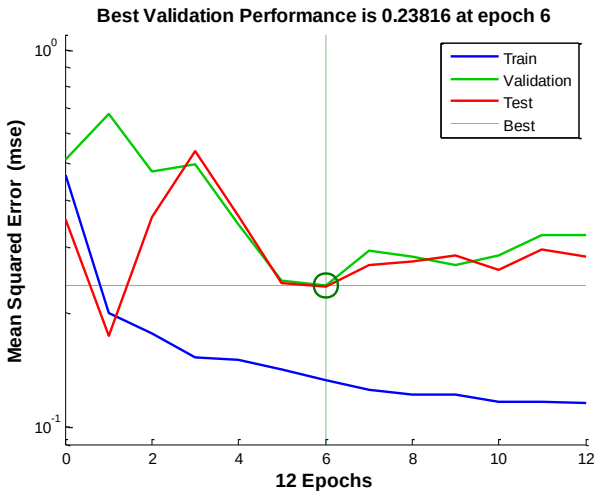

Figure 6.15: Class NED best validation performance

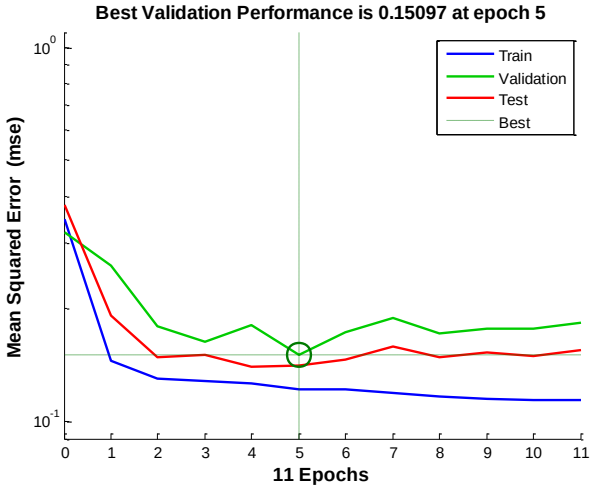

Figure 6.16: Class DOD best validation performance

e) Analysis and comparison of results.

Table 6.20 shows a comparison of performance between the different data integration approaches presented in this chapter. The RBF-NF models (Sanchez-Carbayo NN Input-Output mapping and Sanchez-Carbayo Discretisation) exhibit a higher performance (AUC of the ROC curve) in the three cohorts as compared to the results for cross-validation, median adjust and discretisation.

Table 6.20: Comparison of results between RBF-NF models with 5 rules and 25 inputs

	Testing		
Bladder Cancer	**AUC Sanchez**	**AUC Blaveri**	**AUC Kim**
Performance Cross-validation	0.81	0.45	0.49
Performance Sanchez-Carbayo Median Adjust	0.81	0.53	0.66
Performance Sanchez-Carbayo Discretisation	0.75	0.60	0.60
Performance Sanchez-Carbayo NN mapping	0.81	0.86	0.78

6.5 Summary

In this Chapter, the generalisation performance of the developed models is investigated. The question that investigated in this chapter is why a model that can predict with good accuracy in the same cohort is bad when it is tested on a different data set (cross-validated). Is this challenge arising due to characteristic intrinsic of the data? The approach studied in this chapter is to cross-validate distinct microarray data by applying data integration techniques. Three different data integration approaches are analysed: quantile discretisation, median adjust and NN input-output mapping. The latter two approaches are introduced for the first time to a bladder cancer classification

algorithm. The results obtained demonstrate that the data integration methods for cross validation of the models helps to have a considerable increase in the accuracy.

The first challenge arises because the top genes selected by the classifier do not exist in the different data sets. The common genes between the data sets tend to be a much smaller subset of the original cohort; typically reduced from ten or twenty thousand to a couple of thousand.

The approach studied in this chapter is to cross-validate distinct microarray data by applying data integration techniques. The main challenge is that researchers use different microarray platforms and pre-processing algorithms, making difficult to validate the results found on each data study [174]. Three different data integration approaches are analysed: quantile discretisation, median adjust and NN input-output mapping. The last two approaches are introduced for the first time to a bladder cancer classification algorithm.

The results obtained (Tables 6.13-6.19) demonstrate that the data integration methods for cross validation of the models give an increase in the performance. If the results from the data integration methods are compared to previously published results it can be seen that the NN mapping and Discretisation and median adjust have a higher performance in terms of AUC of a ROC curve. The obtained results demonstrate how data integration methods for model cross-validation can have a significant increase in the generalisation performance, and enable previously developed models to be used in different patient cohorts.

Despite the fact that more information can be extracted from microarray models, the generalisation issue makes them still unreliable for clinicians.

At moment, there is no definitive approach for data integration because most of the results are data-dependent. One of the biggest challenges is that there is no agreement on which pre-processing algorithm should be used to produce comparable expression measurements across different platforms.

Chapter's summary of achievements:

- Improve the generalisation performance in microarray bladder cancer data
- Two different data integration approaches are presented for the first time: median adjust and NN mapping of input-output.
- The results obtained prove that the data integration methods for cross validation of the models helps to have a significant increase in the accuracy.

Chapter 7: Conclusions and future research directions

Through this entire thesis, it has been emphasised why the study of predicting cancer is of great significance. The main reasons are: to help decrease the mortality rate, to make a prompt and correct classification of the type of cancer that would later translate into avoiding unnecessary treatment and save costs. Because traditional prediction tools have struggled to make an accurate classification at the early stages of cancer, new technologies have emerged for the study of cancer. Microarray gene expression data is one of these new technologies. The main challenge that these types of studies run across is the high dimensionality, translated in thousands of genes but a small number of samples. As there are no physics/biology based equations that represent the behaviour of the genes, a predictor model (data-driven) must be produced.

An RBF-NF methodology for the case study of bladder cancer prediction with respect to the patient's stage, grade and survival was proposed. RBF-Neural-Fuzzy models offer balance of performance and simplicity (while being tolerant to imprecision); these are traits that are important in healthcare informatics. The focus of this research is to produce a model to identify the parameters significant to the process (genes) maintain *simplicity* and *transparency* while at the same time makes an accurate prediction of cancer survival. The major benefit of this approach, apart from its good

accuracy, is the transparency given by the rule base, converting the rules from the model into a graphical output that can be better understood in a visual manner. Such traits can aid the development of easy to understand and use model by non-experts (non-engineers) such as clinicians in order to directly interrogate the resulting model (human-centric system).

Compared to previous modelling attempts from Martin Lauss [113] and Riester [114] based on SVM, the developed RBF-NF method shows improved performance in the same datasets. However, the attractiveness of this method is on the transparency that the rule-base exhibits and the good generalisation performance (even with just 20 genes and 5 rules) as compared to previous modelling attempts on the same dataset. The rule-base's transparency and interpretability, can aid the clinicians to directly interrogate the resulting model (human-centric system) and examine how the model uses individual genes and their intensity to provide predictions on the stage, grade and survival of bladder cancer.

The scaling-up performance of Radial Basis Function (RBF) Neural-Fuzzy models is also investigated. The aim was to find the rational limit for the maximum number of useful inputs (genes) to use in the model while still maintaining low computational complexity and high accuracy. An enhanced rule-base extraction framework is proposed to improve the model's performance for high-dimensional low sample size data.

From the results obtained it can be concluded that the RBF model using FCM alone performs best when less than 300 genes are used. Due to the characteristics of high-dimension low sample size data, as the number of genes increases but number of samples remains the same, the WFCM and WFCM with the validity index are needed to model the microarray data with a good level of accuracy.

The developed models maintain the simple structure with just five (5) rules, but with very good performance (up to 2000 genes). The training time for the models can still be up to 3-4 days on a high performance computing server; however other more efficient-optimisation algorithm can be used instead.

One of the main contributions of this research is the introduction of a new input selection method; this new method is based on the polynomial output of the RBF-NF model. The hypothesis behind the new Input selection is to monitor the values of the output weights and membership degree during the training of the structure. Because of the polynomial output of the model, it is conceivable to distinguish how much a gene is involved in the final output and if that rule is important for the system.

The new feature selection algorithm is based on Fuzzy entropy and a RBF Neural-Fuzzy structure that links directly the fuzzy entropy to the relative significance of the features of the model. Because of the characteristics of the RBF-NF TSK output (input weighted polynomial) it is possible to correlate the features that are more significant to the model's prediction. This significance measure is used to rank the inputs of the model via an iterative algorithm.

Another contribution of this research is how the combination of clinical data (stage and grade) as additional inputs to the most commonly used microarray gene intensities improves the overall performance (with various levels of improvement). Crucially, the combination of Stage and Grade and the low number of genes resulted from the approach helps the model to be developed in simpler structures (low number of rules and genes, thus reducing model complexity), while maintaining comparable or improved performance as compared to models with significantly more genes or more complex structure. The combination of Stage and Grade also helps the model to reduce

the training iterations (easier to optimise), helping to reduce the computational cost to just a few seconds on a standard single personal computer.

The biggest challenge though is presented in the generalisation ability of such data-driven models as identified by other research results too. Models that are trained based on a specific patient cohort should be tested against data from other cohorts to establish the developed models' generalisation performance and predictive robustness. The possibility of creating a general model that can be used with any type of microarray data set and still make a prediction with respectable accuracy (around 75%) was also investigated.

The first challenge arises because the top genes selected by the classifier do not exist in the different data sets. The common genes between the data sets tend to be a much smaller subset of the original cohort; typically reduced from ten or twenty thousand to a couple of thousand. The main challenge is that researchers use different microarray platforms and pre-processing algorithms making difficult to validate the results found on each data study [174]. Three different data integration approaches are analysed: quantile discretisation, median adjust and NN input-output mapping. The latter two approaches are introduced for the first time to a bladder cancer classification algorithm.

The results obtained demonstrate that the data integration methods for cross validation of the models give an increase in the performance. If the results from the data integration methods are compared to previously published results it can be seen that the NN mapping and Discretisation have a higher performance in terms of AUC of a ROC curve. The obtained results demonstrate how data integration methods for model cross-validation can have a significant increase in the generalisation performance, and enable

previously developed models to be used in different patient cohorts. Despite the fact that more information can be extracted from microarray models, the generalisation issue makes them still unreliable for clinicians.

The generalisation performance of the predictive methods is the main limitation of this study.

The limitations of this study are given by the nature of microarray data: missing values, noise or error from scanners. The results obtained from this study are data-dependent and are closely related to the quality of the microarray data. It must not be forgotten that the different analysis techniques applied in this study are not a remedy for low quality data.

7.1 Future research directions

The work conducted revealed a number of weaknesses of existing methodologies (hence engineering challenges) for the reliable prediction of cancer from the clustering methods, input selection or the model selected to make the prediction.

As explained in Chapter 2 of this Thesis, every aspect of the a data-driven modelling approach is important and in the past years it was discovered how normalisation can affect significantly the data, the number of inputs a method can work with (complexity dependant) and asses predictive performance via a Neural-Fuzzy approach. Even though the presented methodology is produced via the use of microarray bladder cancer data as a case study, this method may also be applied to numerous other diseases.

The aim of future work might consist of using the developed input selection and new modelling techniques to construct a multidimensional Patient Prognostic Map. This ‘map’ will be a framework that uses the developed hybrid model, the gene selection,

and expert knowledge to provide to the clinicians linguistic advice on cancer progression.

7.1.1 Future research directions for the RBF NF model

i. Fusion the microarray data with clinical screening data as well as temporal data (hybrid model)

The aim is to integrate the microarray data with clinical screening data and Temporal Data (hybrid model). The term Temporal Data refers to a new design of experiments, in which a number of individuals of various cell classes are involved and gene expression is measured for each individual during a time course. In other words, the new experiments measure temporal gene expression multiple times for each cell class, but each time the measurement is performed on different individuals of that cell class.

As described above a hybrid model will be built based on the combination of the data sets, however expert knowledge will need to be embedded into the system to allow the fusion of the two sources of information. Furthermore, Fuzzy Fusion approach will be used to amalgamate all the information produced by the modelling scheme (patient map) and present this in a linguistic form.

ii. Use a diagnosis of stage/grade/survival and based on that diagnosis to inform a treatment therapy

Based in the Predictions of survival, a Medical Diagnosis tool can be developed to assist therapy and treatment for the disease.

The prediction of the malignancy of the cancer will help to avoid unnecessary surgery, improving the life quality of the patient by only receiving absolutely necessary treatment for the disease. Another important improvement is the reduction in the overall therapy costs.

Figure 7.1 shows the overall architecture of the Multidimensional Patient Prognostic Map.

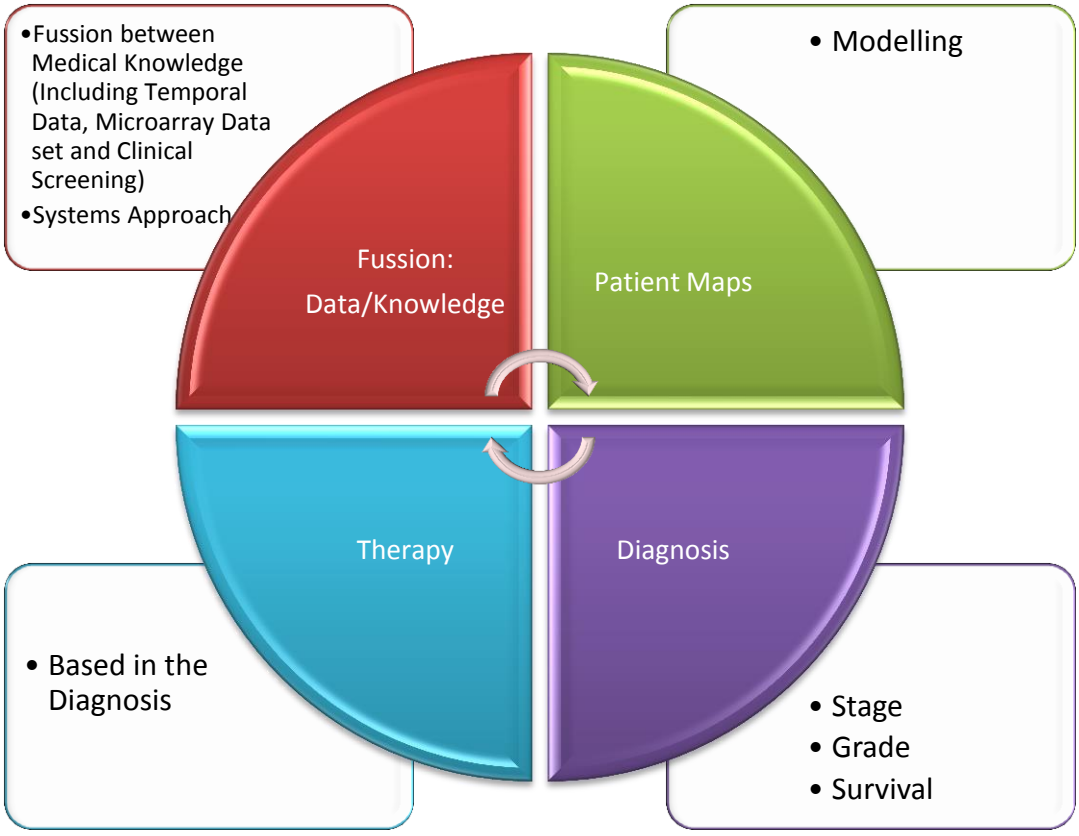

Figure 7.1: Multidimensional Patient Prognostic Maps

In Figure 7.1, a diagram that represents the model is shown. The data to analyse comes from the fusion of Microarrays data set, clinical screening and Temporal Data. The fusion is possible thanks to the Medical knowledge and the Systems approach work. The role of the Medical expertise is going to be needed to help us amalgamate the medical screenings with the data sets and based on that fusion develop the Patient Maps. The proposed model is going to predict Stage, Grade and Survival based on that prediction a Clinical Diagnosis and Therapy will be develop.

7.1.2 Future research directions for microarray analysis

The generalisation performance of new predictive methods needs to be studied. As discussed in this Thesis, the real test for most data-driven cancer prediction models is the test of generalisation, i.e. when the model is confronted with a new patient cohort. The healthcare professionals community (medical, biology, chemistry, and engineering) is required to work together to produce a standard that unifies the analysis of tissue samples, image processing, normalisation and representation of gene intensities.

There are problems with microarray that perhaps will never get solved; however there is still place for improvement, for example:

- Noise or error from scanners need to be reduced to the minimum
- Reduction of the number of missing gene expression values

Because microarray analysis has been conducted since several years ago, it is important to *re-use* different studies (data sets) and validate the results previously obtained. It is also important to continue the research in the area of meta-analysis and data integration methods towards robust classification results.

References

[1] C. J. L. Murray and A. D. Lopez, "Alternative projections of mortality and disability by cause 1990–2020: Global Burden of Disease Study," *The Lancet*, vol. 349, pp. 1498-1504, 1997.

[2] R. A. Weinberg, *The biology of cancer*. New York: Garland Science, 2007.

[3] E. B. Avritscher, C. D. Cooksley, H. B. Grossman, A. L. Sabichi, L. Hamblin, C. P. Dinney, and L. S. Elting, "Clinical model of lifetime cost of treating bladder cancer and associated complications," *Urology*, vol. 68, pp. 549-53, Sep 2006.

[4] H. B. Burke, P. H. Goodman, D. B. Rosen, D. E. Henson, J. N. Weinstein, F. E. Harrell Jr, 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.

[5] D. G. Bostwick and H. B. Burke, "Prediction of individual patient outcome in cancer: comparison of artificial neural networks and Kaplan--Meier methods," *Cancer*, vol. 91, pp. 1643-1646, 2001.

[6] M. H. Te and V. Kecman, "Gene extraction for cancer diagnosis by support vector machines - An improvement," *Artificial Intelligence in Medicine*, vol. 35, pp. 185-194, 2005.

[7] J. W. F. Catto, D. A. Linkens, M. F. Abbod, M. Chen, J. L. Burton, K. M. Feeley, and F. C. Hamdy, "Artificial intelligence in predicting bladder cancer outcome: A comparison of neuro-fuzzy modeling and artificial neural networks," *Clinical Cancer Research*, vol. 9, pp. 4172-4177, 2003.

[8] S. Sun, Q. Peng, and A. Shakoor, "A kernel-based multivariate feature selection method for microarray data classification," *PLoS ONE*, vol. 9, 2014.

[9] E. Rosenberg, J. Baniel, Y. Spector, A. Faerman, E. Meiri, R. Aharonov, D. Margel, Y. Goren, and O. Nativ, "Predicting progression of bladder urothelial carcinoma using microRNA expression," *BJU International*, pp. n/a-n/a, 2013.

[10] A. Zibakhsh and M. S. Abadeh, "Gene selection for cancer tumor detection using a novel memetic algorithm with a multi-view fitness function," *Engineering Applications of Artificial Intelligence*, vol. 26, pp. 1274-1281, 2013.

[11] Y. Zhang, J. Xuan, R. Clarke, and H. W. Ressom, "Module-based breast cancer classification," *International Journal of Data Mining and Bioinformatics*, vol. 7, pp. 284-302, 2013.

[12] C. K. Chen, "The classification of cancer stage microarray data," *Computer Methods and Programs in Biomedicine*, vol. 108, pp. 1070-1077, 2012.

[13] D. Tilki, M. Burger, G. Dalbagni, H. B. Grossman, O. W. Hakenberg, J. Palou, O. Reich, M. Rouprêt, S. F. Shariat, and A. R. Zlotta, "Urine markers for detection and surveillance of non-muscle-invasive bladder cancer," *Eur Urol*, vol. 60, pp. 484-492, 2011.

[14] M. G. W. Bol, J. P. A. Baak, S. Buhr-Wildhagen, A.-J. Kruse, K. H. Kjellevold, E. A. M. Janssen, O. Mestad, and P. E. R. ØGreid, "Reproducibility and Prognostic Variability of Grade and Lamina Propria Invasion in Stages Ta, T1 Urothelial Carcinoma of the Bladder," *The Journal of Urology*, vol. 169, pp. 1291-1294, 2003.

[15] A. Kalousis, J. Prados, and M. Hilario, "Stability of feature selection algorithms: A study on high-dimensional spaces," *Knowledge and Information Systems*, vol. 12, pp. 95-116, 2007.

[16] M. F. Abbod, F. C. Hamdy, D. A. Linkens, and J. W. F. Catto, "Predicted modeling in cancer: Where systems biology meets the stock market," *Expert Review of Anticancer Therapy,* vol. 9, pp. 867-870, 2009.

[17] J. W. F. Catto, M. F. Abbod, D. A. Linkens, and F. C. Hamdy, "Neuro-fuzzy modeling: An accurate and interpretable method for predicting bladder cancer progression," *Journal of Urology,* vol. 175, pp. 474-479, 2006.

[18] J. W. F. Catto, M. F. Abbod, P. J. Wild, D. A. Linkens, C. Pilarsky, I. Rehman, D. J. Rosario, S. Denzinger, M. Burger, R. Stoehr, R. Knuechel, A. Hartmann, and F. C. Hamdy, "The Application of Artificial Intelligence to Microarray Data: Identification of a Novel Gene Signature to Identify Bladder Cancer Progression," *European Urology,* vol. 57, pp. 398-406, 2010.

[19] M. F. Abbod, J. W. F. Catto, D. A. Linkens, P. J. Wild, A. Herr, C. Wissmann, C. Pilarsky, A. Hartmann, and F. C. Hamdy, "Artificial Intelligence Technique for Gene Expression Profiling of Urinary Bladder Cancer," in *Intelligent Systems, 2006 3rd International IEEE Conference on*, 2006, pp. 646-651.

[20] M. Brown, K. M. Bossley, D. J. Mills, and C. J. Harris, "High dimensional neurofuzzy systems: overcoming the curse of dimensionality," in *Fuzzy Systems, 1995. International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium., Proceedings of 1995 IEEE Int*, 1995, pp. 2139-2146 vol.4.

[21] S. Chavan, F. Bray, J. Lortet-Tieulent, M. Goodman, and A. Jemal, "International variations in bladder cancer incidence and mortality," *Eur Urol,* vol. 66, pp. 59-73, Jul 2014.

[22] D. Raghavan and M. Bailey, *Bladder Cancer*: Health, 2006.

[23] C. R. UK. (2014, August 2014). *Cancer Research Satistics*. Available: http://www.cancerresearchuk.org

[24] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, pp. 2507-2517, 2007.

[25] A. Jemal, R. Siegel, E. Ward, Y. Hao, J. Xu, and M. J. Thun, "Cancer Statistics, 2009," *CA: A Cancer Journal for Clinicians*, vol. 59, pp. 225-249, 2009.

[26] N. M. Arts, "Bladder and Nearby Organs (Female)," vol. 818 × 818 (43 KB). http://commons.wikimedia.org/wiki/File:Bladder_and_nearby_organs_(female).jpg: Wikimedia Commons and National Cancer Institute, 2001.

[27] G. B. Di Pierro, C. Gulia, C. Cristini, G. Fraietta, L. Marini, P. Grande, V. Gentile, and R. Piergentili, "Bladder cancer: a simple model becomes complex," *Curr Genomics*, vol. 13, pp. 395-415, Aug 2012.

[28] S. H. Landis, T. Murray, S. Bolden, and P. A. Wingo, "Cancer statistics, 1999," *CA a cancer journal for clinicians*, vol. 49, pp. 8-31, 1, 1999.

[29] F. Thomas, A. P. Noon, N. Rubin, J. Goepel, and J. W. Catto, "Comparative outcomes of primary, recurrent and progressive high-risk non-muscle invasive bladder cancer," *Eur Urol*, vol. 63, pp. 145-54, 2013.

[30] M. Rink, E. Xylinas, V. Margulis, E. K. Cha, B. Ehdaie, J. D. Raman, F. K. Chun, K. Matsumoto, Y. Lotan, H. Furberg, M. Babjuk, A. Pycha, C. G. Wood, P. I. Karakiewicz, M. Fisch, D. S. Scherr, and S. F. Shariat, "Impact of smoking on oncologic outcomes of upper tract urothelial carcinoma after radical nephroureterectomy," *European urology*, vol. 63, pp. 1082-90, Jun 2013.

[31] A. P. Noon, P. C. Albertsen, F. Thomas, D. J. Rosario, and J. W. Catto, "Competing mortality in patients diagnosed with bladder cancer: evidence of undertreatment in the elderly and female patients," *Br J Cancer*, vol. 108, pp. 1534-40, Apr 16 2013.

[32] B. H. Channel. (2014, August 2014). *Bladder Cancer*. Available: http://www.betterhealthchannel.vic.gov.au/bhcv2/bhcpdf.nsf/ByPDF/Bladder_cancer/$File/Bladder_cancer.pdf

[33] R. S. Svatek, B. K. Hollenbeck, S. Holmang, R. Lee, S. Kim, A. Stenzl, and Y. Lotan, "The economics of bladder cancer: Costs and considerations of caring for this disease," *Eur Urol*, 2014.

[34] J. W. Catto, A. Alcaraz, A. S. Bjartell, R. De Vere White, C. P. Evans, S. Fussel, F. C. Hamdy, O. Kallioniemi, L. Mengual, T. Schlomm, and T. Visakorpi, "MicroRNA in Prostate, Bladder, and Kidney Cancer: A Systematic Review," *Eur Urol*, vol. 59, pp. 671-81, Feb 1 2011.

[35] J. W. Catto, S. Miah, H. C. Owen, H. Bryant, E. Dudziec, S. Larre, M. Milo, I. Rehman, D. J. Rosario, E. DiMartino, M. A. Knowles, M. Meuth, A. L. Harris, and F. C. Hamdy, "Distinct microRNA alterations characterize high and low grade bladder cancer," *Cancer Res*, vol. 69, pp. 8472-81, 2009.

[36] A. Hayes and D. C. Hoyle, "Microarray gene expression data analysis: a beginners guide by Helen C. Causton, John Quackenbush and Alvis Brazma. Blackwell Science, Oxford, UK," *Yeast*, vol. 21, pp. 973-974, 2004.

[37] F. Ducray, J. Honnorat, and J. Lachuer, "DNA microarray technology: Principles and applications to the study of neurological disorders," *Principes et intérêts pour l'étude des maladies neurologiques et technologie des puces ADN*, vol. 163, pp. 409-420, 2007.

[38] E. Suárez, A. Burguete, and G. J. McLachlan, "Microarray data analysis for differential expression: A tutorial," *Puerto Rico health sciences journal*, vol. 28, pp. 89-104, 2009.

[39] G. Paumier, "DNA Microarray," vol. 820 × 820 (293 KB), D. Microarray, Ed., ed. http://commons.wikimedia.org/wiki/File:DNA_microarray.svg: Wikimedia Commons, 2008.

[40] M. Ford, *Medical Microbiology*: OUP Oxford, 2014.

[41] G. B. Whitworth, "An introduction to microarray data analysis and visualization," vol. 470, ed, 2010, pp. 19-50.

[42] I. Guyon, Andr, #233, and Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res.*, vol. 3, pp. 1157-1182, 2003.

[43] I. Inza, P. Larrañaga, R. Blanco, and A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domains," *Artificial Intelligence in Medicine*, vol. 31, pp. 91-103, 2004.

[44] V. Bolón-Canedo, N. Sánchez-Maroño, A. Alonso-Betanzos, J. M. Benítez, and F. Herrera, "A review of microarray datasets and applied feature selection methods," *Information Sciences*, vol. 282, pp. 111-135, 2014.

[45] P. N. Tan, M. Steinbach, and V. Kumar, "Cluster Analysis: Basic Concepts and Algorithms," *Introduction to Data Mining*, 2005.

[46] P. Jafari and F. Azuaje, "An assessment of recently published gene expression data analyses: Reporting experimental design and statistical factors," *BMC Medical Informatics and Decision Making*, vol. 6, 2006.

[47] M. A. Hall and L. A. Smith, "Practical feature subset selection for machine learning," 1998.

[48] M. Hall, "Correlation-Based Feature Selection for Machine Learning.," Doctor of Philosophy, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1999.

[49] L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution," in *Proceedings, Twentieth International Conference on Machine Learning*, Washington, DC, 2003, pp. 856-863.

[50] J. Wang, L. Wu, J. Kong, Y. Li, and B. Zhang, "Maximum weight and minimum redundancy: A novel framework for feature subset selection," *Pattern Recognition,* vol. 46, pp. 1616-1627, 2013.

[51] V. Bolón-Canedo, S. Seth, N. Sánchez-Maroño, A. Alonso-Betanzos, and J. C. Príncipe, "Statistical dependence measure for feature selection in microarray datasets," in *ESANN 2011 proceedings, 19th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, 2010, pp. 23-28.

[52] P. E. Meyer, C. Schretter, and G. Bontempi, "Information-Theoretic Feature Selection in Microarray Data Using Variable Complementarity," *Selected Topics in Signal Processing, IEEE Journal of,* vol. 2, pp. 261-274, 2008.

[53] F. F. G. Navarro and L. A. B. Muñoz, "Gene subset selection in microarray data using entropic filtering for cancer classification," *Expert Systems,* vol. 26, pp. 113-124, 2009.

[54] A. J. Ferreira and M. A. T. Figueiredo, "An unsupervised approach to feature discretization and selection," *Pattern Recognition,* vol. 45, pp. 3048-3060, 9// 2012.

[55] V. Bolon-Canedo, N. Sanchez-Marono, and A. Alonso-Betanzos, "On the effectiveness of discretization on gene selection of microarray data," in *Neural Networks (IJCNN), The 2010 International Joint Conference on*, 2010, pp. 1-8.

[56] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artificial Intelligence,* vol. 97, pp. 273-324, 1997.

[57] H. Liu and H. Motoda, *Feature Selection for Knowledge Discovery and Data Mining*: Kluwer Academic Publishers, 1998.

[58] I. Inza, B. Sierra, R. Blanco, and P. Larrañaga, "Gene selection by sequential search wrapper approaches in microarray cancer class prediction," *Journal of Intelligent and Fuzzy Systems,* vol. 12, pp. 25-33, 2002.

[59] A. Sharma, S. Imoto, and S. Miyano, "A top-r feature selection algorithm for microarray gene expression data," *IEEE/ACM Transactions on Computational Biology and Bioinformatics,* vol. 9, pp. 754-764, 2012.

[60] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, "Gene selection for cancer classification using support vector machines," *Machine Learning,* vol. 46, pp. 389-422, 2002.

[61] S. Maldonado, R. Weber, and J. Basak, "Simultaneous feature selection and classification using kernel-penalized support vector machines," *Information Sciences,* vol. 181, pp. 115-128, 2011.

[62] W. Awada, T. M. Khoshgoftaar, D. Dittman, R. Wald, and A. Napolitano, "A review of the stability of feature selection techniques for bioinformatics data," in *2012 IEEE 13th International Conference on Information Reuse and Integration, IRI 2012*, Las Vegas, NV, 2012, pp. 356-363.

[63] C. Lazar, J. Taminau, S. Meganck, D. Steenhoff, A. Coletta, C. Molter, V. de Schaetzen, R. Duque, H. Bersini, and A. Nowe, "A Survey on Filter Techniques for Feature Selection in Gene Expression Microarray Analysis," *Computational Biology and Bioinformatics, IEEE/ACM Transactions on,* vol. 9, pp. 1106-1119, 2012.

[64] G. Forman, "An extensive empirical study of feature selection metrics for text classification," *J. Mach. Learn. Res.*, vol. 3, pp. 1289-1305, 2003.

[65] A. P. Engelbrecht, *Computational Intelligence: An Introduction*: Wiley Publishing, 2007.

[66] B. Karlik, "SOFT COMPUTING METHODS IN BIOINFORMATICS: A COMPREHENSIVE REVIEW," *Mathematical and Computational Applications,* vol. 18, pp. 176-197, 2013.

[67] L. A. Zadeh, "Fuzzy logic, neural networks, and soft computing," *Commun. ACM,* vol. 37, pp. 77-84, 1994.

[68] Z. Wang and V. Palade, "Building interpretable fuzzy models for high dimensional data analysis in cancer diagnosis," *BMC Genomics,* vol. 12, p. S5, 2011.

[69] G. Schaefer, T. Nakashima, and Y. Yokota, "Fuzzy Classification for Gene Expression Data Analysis," *Computational Intelligence in Bioinformatics,* pp. 209 - 218, 2008.

[70] E. B. Huerta, B. Duval, and J.-K. Hao, "Fuzzy logic for elimination of redundant information of microarray data," *Genomics proteomics bioinformatics Beijing Genomics Institute,* vol. 6, pp. 61-73, 2008.

[71] P. Woolf, Y. Wang, and P. J, "A Fuzzy Logic Approach to Analyzing Gene Expression Data," *Physiol Genomics,* vol. 3, pp. 9 - 15, 2000.

[72] Z. Wang, V. Palade, and Y. Xu, "Neuro-Fuzzy Ensemble Approach for Microarray Cancer Gene Expression Data Analysis," *Proc of the Second International Symposium on Evolving Fuzzy System EFS06 IEEE Computational Intelligence Society,* pp. 241-246, 2006.

[73] M. Xiong, W. Li, J. Zhao, L. Jin, and E. Boerwinkle, "Feature (gene) selection in gene expression-based tumor classification," *Molecular Genetics and Metabolism,* vol. 73, pp. 239-247, 2001.

[74] L. J. Lancashire, C. Lemetre, and G. R. Ball, "An introduction to artificial neural networks in bioinformatics--application to complex microarray and mass spectrometry datasets in cancer studies," *BRIEFINGS IN BIOINFORMATICS*, vol. 10, pp. 315-329, 2009.

[75] H. T. Huynh, J. J. Kim, and Y. Won, "Classification study on DNA microarray with feedforward neural network trained by singular value decomposition," *International Journal of Bio-Science and Bio-Technology*, vol. 1, pp. 17-24, 2009.

[76] M. C. O. N. a. L. Song, "Neural network analysis of lymphoma microarray data: prognosis and diagnosis near-perfect," *BMC Bioinformatics*, 2003.

[77] J. Khan, J. S. Wei, M. Ringnér, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, C. Peterson, and P. S. Meltzer, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, pp. 673-679, 2001.

[78] L. J. Lancashire, D. G. Powe, J. S. Reis-Filho, E. Rakha, C. Lemetre, B. Weigelt, T. M. Abdel-Fatah, A. R. Green, R. Mukta, R. Blamey, E. C. Paish, R. C. Rees, I. O. Ellis, and G. R. Ball, "A validated gene expression profile for detecting clinical outcome in breast cancer using artificial neural networks," *Breast Cancer Research and Treatment*, vol. 120, pp. 83-93, 2010.

[79] J.-S. R. Jang and C.-T. Sun, "Neuro-fuzzy modeling and control," *Proceedings of the IEEE*, vol. 83, pp. 378-406, 1995.

[80] S. Mitra and Y. Hayashi, "Neuro-fuzzy rule generation: survey in soft computing framework," *IEEE Transactions on Neural Networks*, vol. 11, pp. 748-768, 2000.

[81] P. C. Nayak, K. P. Sudheer, D. M. Rangan, and K. S. Ramasastri, "A neuro-fuzzy computing technique for modeling hydrological time series," *Journal of Hydrology,* vol. 291, pp. 52-66, 2004.

[82] W. Q. Wang, M. F. Golnaraghi, and F. Ismail, "Prognosis of machine health condition using neuro-fuzzy systems," *Mechanical Systems and Signal Processing,* vol. 18, pp. 813-831, 2004.

[83] L. Zadeh, "Fuzzy Sets," *Information and Control,* 1965.

[84] L. A. Zadeh, "Soft computing and fuzzy logic," *IEEE Software,* vol. 11, pp. 48-56, 1994.

[85] N. N. Karnik, J. M. Mendel, and Q. Liang, "Type-2 fuzzy logic systems," *IEEE Transactions on Fuzzy Systems,* vol. 7, pp. 643-658, 1999.

[86] V. Adriaenssens, B. De Baets, P. L. M. Goethals, and N. De Pauw, "Fuzzy rule-based models for decision support in ecosystem management," *Science of the Total Environment,* vol. 319, pp. 1-12, 2004.

[87] S. M. Samuri, G. Panoutsos, M. Mahfouf, G. H. Mills, M. Denaï, and B. H. Brown, "Neural-fuzzy modelling of lung volume using absolute electrical impedance tomography," in *International Conference on Bio-Inspired Systems and Signal Processing, BIOSIGNALS*, Rome, 2011, pp. 43-50.

[88] C. S. Nunes, M. Mahfouf, D. A. Linkens, and J. E. Peacock, "Modelling and multivariable control in anaesthesia using neural-fuzzy paradigms: Part I. Classification of depth of anaesthesia and development of a patient model," *Artificial Intelligence in Medicine,* vol. 35, pp. 195-206, 2005.

[89] M. Denaï, M. Mahfouf, O. K. King, and J. J. Ross, "Online qualitative abstraction of cardiovascular hemodynamics for post cardiac surgery decision support," Bethesda, MD, 2008, pp. 47-50.

[90] B. Schölkopf and A. J. Smola, *Learning with Kernels* vol. 64: MIT Press, 2002.

[91] M. P. S. Brown, W. N. Grundy, D. Lin, N. Cristianini, C. W. Sugnet, T. S. Furey, M. Ares Jr, and D. Haussler, "Knowledge-based analysis of microarray gene expression data by using support vector machines," *Proceedings of the National Academy of Sciences of the United States of America,* vol. 97, pp. 262-267, 2000.

[92] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference and prediction*: Springer, 2009.

[93] J. A. K. Suykens, T. Van Gestel, J. Vandewalle, and B. De Moor, "A support vector machine formulation to PCA analysis and its kernel version," *IEEE Transactions on Neural Networks,* vol. 14, pp. 447-450, 2003.

[94] M. Hecker, S. Lambeck, S. Toepfer, E. Van Someren, and R. Guthke, "Gene regulatory network inference: data integration in dynamic models-a review," *Bio Systems,* vol. 96, pp. 86-103, 2009.

[95] A. V. Werhli and D. Husmeier, "Reconstructing gene regulatory networks with bayesian networks by combining expression data with multiple sources of prior knowledge," *Statistical Applications in Genetics and Molecular Biology,* vol. 6, pp. Article15-Article15, 2007.

[96] M. Korucuoglu, S. Isci, A. Ozgur, and H. H. Otu, "Bayesian pathway analysis of cancer microarray data," *PLoS ONE,* vol. 9, p. e102803, 2014.

[97] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of computational biology a journal of computational molecular cell biology,* vol. 7, pp. 601-620, 2000.

[98] A. Polanski, J. Polanska, M. Jarzab, M. Wiench, and B. Jarzab, "Application of Bayesian networks for inferring cause-effect relations from gene expression profiles of cancer versus normal cells," *Mathematical Biosciences,* vol. 209, pp. 528-546, 2007.

[99] F. K. Ahmad, S. Deris, and N. H. Othman, "The inference of breast cancer metastasis through gene regulatory networks," *J Biomed Inform*, vol. 45, pp. 350-62, Apr 2012.

[100] S. Chakraborty and M. Ghosh, "Applications of Bayesian Neural Networks in Prostate Cancer Study," vol. 28, ed, 2012, pp. 241-262.

[101] I. Bichindaritz and A. Annest, *Case based reasoning with Bayesian model averaging: An improved method for survival analysis on microarray data* vol. 6176 LNAI. Alessandria, 2010.

[102] N. Friedman, "The Bayesian structural EM algorithm," in *In UAI*, 1998, pp. 129-138.

[103] D. M. Chickering, "Learning Equivalence Classes of Bayesian-Network Structures," *Journal of Machine Learning Research*, vol. 2, pp. 445-498, 2002.

[104] L. Mengual, M. Burset, E. Ars, J. J. Lozano, H. Villavicencio, M. J. Ribal, and A. Alcaraz, "DNA Microarray Expression Profiling of Bladder Cancer Allows Identification of Noninvasive Diagnostic Markers," *Journal of Urology*, vol. 182, pp. 741-748, 2009.

[105] S. Egawa and H. Kuruma, "Search for Biomarkers of Aggressiveness in Bladder Cancer," *Eur Urol*, vol. 50, pp. 20-22, 2006.

[106] M. Sanchez-Carbayo, N. D. Socci, J. Lozano, F. Saint, and C. Cordon-Cardo, "Defining molecular profiles of poor outcome in patients with invasive bladder cancer using oligonucleotide microarrays," *Journal of Clinical Oncology*, vol. 24, pp. 778-789, 2006.

[107] W. J. Kim, E. J. Kim, S. K. Kim, Y. J. Kim, Y. S. Ha, P. P. Jeong, M. J. Kim, S. J. Yun, K. M. Lee, S. K. Moon, S. C. Lee, E. J. Cha, and S. C. Bae, "Predictive value of progression-related gene classifier in primary non-muscle invasive bladder cancer," *Molecular Cancer*, vol. 9, 2010.

[108] S. F. Shariat, P. I. Karakiewicz, R. Ashfaq, S. P. Lerner, G. S. Palapattu, R. J. Cote, A. I. Sagalowsky, and Y. Lotan, "Multiple biomarkers improve prediction of bladder cancer recurrence and mortality in patients undergoing cystectomy," *Cancer,* vol. 112, pp. 315-325, 2008.

[109] E. Dudziec, A. Gogol-Doring, V. Cookson, W. Chen, and J. Catto, "Integrated epigenome profiling of repressive histone modifications, DNA methylation and gene expression in normal and malignant urothelial cells," *PLoS One,* vol. 7, p. e32750, 2012.

[110] E. Blaveri, J. P. Simko, J. E. Korkola, J. L. Brewer, F. Baehner, K. Mehta, S. DeVries, T. Koppie, S. Pejavar, P. Carroll, and F. M. Waldman, "Bladder cancer outcome and subtype classification by gene expression," *Clinical Cancer Research,* vol. 11, pp. 4044-4055, 2005.

[111] L. Dyrskjøt, T. Thykjaer, M. Kruhøffer, J. L. Jensen, N. Marcussen, S. Hamilton-dutoit, H. Wolf, and T. F. Ørntoft, "Identifying distinct classes of bladder carcinoma using microarrays," *Online,* vol. 33, pp. 90-96, 2003.

[112] M. Sanchez-Carbayo, N. D. Socci, J. J. Lozano, W. Li, E. Charytonowicz, T. J. Belbin, M. B. Prystowsky, A. R. Ortiz, G. Childs, and C. Cordon-Cardo, "Gene discovery in bladder cancer progression using cDNA microarrays," *American Journal of Pathology,* vol. 163, pp. 505-516, 2003.

[113] M. Lauss, M. Ringnér, and M. Höglund, "Prediction of stage, grade, and survival in bladder cancer using genome-wide expression data: a validation study," *Clinical Cancer Research,* vol. 16, pp. 4421-4433, 2010.

[114] M. Riester, J. M. Taylor, A. Feifer, T. Koppie, J. E. Rosenberg, R. J. Downey, B. H. Bochner, and F. Michor, "Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer," *Clinical Cancer Research,* vol. 18, pp. 1323-1333, 2012.

[115] B. H. Bochner, G. Dalbagni, M. W. Kattan, P. Fearn, K. Vora, S. S. Hee, L. Zoref, H. Abol-Enein, M. A. Ghoneim, P. T. Scardino, D. Bajorin, D. G. Skinner, J. P. Stein, G. Miranda, J. E. Gschwend, B. G. Volkmer, R. E. Hautmann, S. Chang, M. Cookson, J. A. Smith, G. Thalman, U. E. Studer, C. T. Lee, J. Montie, D. Wood, J. Palou, Y. Fradet, L. LaCombe, P. Simard, M. P. Schoenberg, S. Lerner, A. Vazina, P. Bassi, M. Murai, and E. Kikuchi, "Postoperative nomogram predicting risk of recurrence after radical cystectomy for bladder cancer," *Journal of Clinical Oncology*, vol. 24, pp. 3967-3972, 2006.

[116] S. Chakraborty and M. Ghosh, "Applications of Bayesian Neural Networks in Prostate Cancer Study," vol. 28, ed, 2012, pp. 241-262.

[117] M. G. Schrauder, R. Strick, R. Schulz-Wendtland, P. L. Strissel, L. Kahmann, C. R. Loehberg, M. P. Lux, S. M. Jud, A. Hartmann, A. Hein, C. M. Bayer, M. R. Bani, S. Richter, B. R. Adamietz, E. Wenkel, C. Rauh, M. W. Beckmann, and P. A. Fasching, "Circulating micro-RNAs as potential blood-based markers for early stage breast cancer detection," *PLoS ONE*, vol. 7, 2012.

[118] M. Bremer, E. Himelblau, and A. Madlung, "Introduction to the statistical analysis of two-color microarray data," *Methods Mol Biol*, vol. 620, pp. 287 - 313, 2010.

[119] A. Keller, P. Leidinger, J. Lange, A. Borries, H. Schroers, M. Scheffler, H. P. Lenhof, K. Ruprecht, and E. Meese, "Multiple Sclerosis: MicroRNA Expression Profiles Accurately Differentiate Patients with Relapsing-Remitting Disease from Healthy Controls," *PLoS ONE*, vol. 4, 2009.

[120] T. Hanai, H. Hamada, and M. Okamoto, "Application of bioinformatics for DNA microarray data to bioscience, bioengineering and medical fields," *J Biosci Bioeng*, vol. 101, pp. 377-84, 2006.

[121] A. Szabo, K. Boucher, W. Carroll, L. Klebanov, A. Tsodikov, and A. Yakovlev, "Variable selection and pattern recognition with gene expression data generated by the microarray technology," *Math Biosci,* vol. 176, pp. 71 - 98, 2002.

[122] J. G. Moreno-Torres, J. A. Saez, and F. Herrera, "Study on the impact of partition-induced dataset shift on k-fold cross-validation," *IEEE Trans Neural Netw Learn Syst,* vol. 23, pp. 1304-12, Aug 2012.

[123] P. K. P, P. Vadakkepat, and L. A. Poh, "Fuzzy-rough discriminative feature selection and classification algorithm, with application to microarray and image datasets," *Applied Soft Computing,* vol. 11, pp. 3429-3440, 2011.

[124] B. Efron, "Bootstrap Methods: Another Look at the Jackknife," *The Annals of Statistics,* vol. 7, pp. 1-26, 1979.

[125] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Transactions on Systems, Man and Cybernetics,* vol. 15, pp. 116-132, 1985.

[126] L. Wen and Y. Hori, "An Algorithm for Extracting Fuzzy Rules Based on RBF Neural Network," *Industrial Electronics, IEEE Transactions on,* vol. 53, pp. 1269-1276, 2006.

[127] M. Reimers, "Making Informed Choices about Microarray Data Analysis," *PLoS Computational Biology,* vol. 6, p. e1000786, 2010.

[128] X. Qiu, H. Wu, and R. Hu, "The impact of quantile and rank normalization procedures on the testing power of gene differential expression analysis," *BMC Bioinformatics,* vol. 14, p. 124, 2013.

[129] M. Robinson and T. Speed, "A comparison of Affymetrix gene expression arrays," *BMC Bioinformatics,* vol. 8, p. 449, 2007.

[130] R. Schmid, P. Baum, C. Ittrich, K. Fundel-Clemens, W. Huber, B. Brors, R. Eils, A. Weith, D. Mennerich, and K. Quast, "Comparison of normalization methods

for Illumina BeadChip HumanHT-12 v3," *BMC Genomics,* vol. 11, p. 349, 2010.

[131] G. R. C., C. V. J., B. D. M., B. B., D. M., D. S., E. B., G. L., G. Y., G. J., H. K., H. T., H. W., I. S., I. R., L. F., L. C., M. M., R. A. J., S. G., S. C., S. G., T. L., Y. J. Y., and Z. J., "- Bioconductor: open software development for computational biology and," *Genome Biol. 2004;5(10):R80. Epub 2004 Sep 15.*

[132] L. T. Payam Refaelizadeh, Huan Liu, "Cross-Validation," 2008.

[133] V. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proc Natl Acad Sci USA,* vol. 98, pp. 5116 - 5121, 2001.

[134] T. Dietterich, "Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms," *Neural computation,* vol. 10, pp. 1895-1923, 1998.

[135] J. C. Dunn, "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters," *Cybernetics and Systems,* vol. 3, pp. 32-57, 1973.

[136] J. C. Bezdek, "Pattern Recognition with Fuzzy Objective Function," *Plenum Press,* 1981.

[137] J. Wang, T. H. Bø, I. Jonassen, O. Myklebost, and E. Hovig, "Tumor classification and marker gene prediction by feature selection and fuzzy c-means clustering using microarray data," *BMC Bioinformatics,* vol. 4, 2003.

[138] S. Bandyopadhyay, A. Mukhopadhyay, and U. Maulik, "An improved algorithm for clustering gene expression data," *BIOINFORMATICS,* vol. 23, pp. 2859-2865, 2007.

[139] G. Panoutsos and M. Mahfouf, "A neural-fuzzy modelling framework based on granular computing: Concepts and applications," *Fuzzy Sets and Systems*, vol. 161, pp. 2808-2830, 2010.

[140] U. H. Mazlan and P. Saad, "Classification of breast cancer microarray data using radial basis function network," Langkawi, Kedah, 2012, pp. 41-44.

[141] H. Q. Wang and D. S. Huang, "Non-linear cancer classification using a modified radial basis function classification algorithm," *Journal of Biomedical Science*, vol. 12, pp. 819-826, 2005.

[142] A. Keller, P. Leidinger, A. Borries, A. Wendschlag, F. Wucherpfennig, M. Scheffler, H. Huwer, H. Lenhof, and E. Meese, "MiRNAs in lung cancer - Studying complex fingerprints in patient's blood cells by microarray experiments," *BMC Cancer*, vol. 9, p. 353, 2009.

[143] M. I. A. Lourakis, *{A brief description of the Levenberg-Marquardt algorithm implemented by levmar}*, 2005.

[144] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *The Quarterly of Applied Mathematics*, vol. 2, pp. 164-168, 1944.

[145] D. W. Marquardt, "An Algorithm for Least-Squares Estimation of Nonlinear Parameters," *Journal of the Society for Industrial and Applied Mathematics*, vol. 11, pp. 431-441, 1963.

[146] H. D. Mittelmann, "The Least Squares Problem," 2004.

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

[148] E. Bonilla Huerta, J. C. Hernández Hernández, and L. A. Hernández Montiel, "A New Combined Filter-Wrapper Framework for Gene Subset Selection with Specialized Genetic Operators," in *Advances in Pattern Recognition*. vol. 6256,

J. Martínez-Trinidad, J. Carrasco-Ochoa, and J. Kittler, Eds., ed: Springer Berlin Heidelberg, 2010, pp. 250-259.

[149] T. Li, C. Zhang, and M. Ogihara, "A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression," *Bioinformatics*, vol. 20, pp. 2429-2437, 2004.

[150] C. D. Hong Chai, "An Evaluation of Gene Selection Methods for Multi-class Microarray Data Classification," *In Proceeding of the workshop W9 on data mining and text mining for bioinformatics*, pp. 3-10, 2004.

[151] H. A. Koepke and B. S. Clarke, "On the limits of clustering in high dimensions via cost functions," *Stat. Anal. Data Min.*, vol. 4, pp. 30-53, 2011.

[152] M. Ramze Rezaee, B. P. F. Lelieveldt, and J. H. C. Reiber, "A new cluster validity index for the fuzzy c-mean," *Pattern Recognition Letters*, vol. 19, pp. 237-246, 1998.

[153] G. v. Borries and H. Wang, "Partition clustering of high dimensional low sample size data based on p-values," *Comput. Stat. Data Anal.*, vol. 53, pp. 3987-3998, 2009.

[154] S. Tavazoie, J. D. Hughes, M. J. Campbell, R. J. Cho, and G. M. Church, "Systematic determination of genetic network architecture," *Nat Genet*, vol. 22, pp. 281-5, 1999.

[155] A. D. Niros and G. E. Tsekouras, "On training radial basis function neural networks using optimal fuzzy clustering," in *Control and Automation, 2009. MED '09. 17th Mediterranean Conference on*, 2009, pp. 395-400.

[156] T. C. Havens, J. C. Bezdek, C. Leckie, L. O. Hall, and M. Palaniswami, "Fuzzy c-Means algorithms for very large data," *IEEE Transactions on Fuzzy Systems*, vol. 20, pp. 1130-1146, 2012.

[157] J. C. Bezdek, "Numerical taxonomy with fuzzy sets," *Journal of Mathematical Biology*, vol. 1, pp. 57-71, 1974/05/01 1974.

[158] J. C. Bezdek, "Cluster Validity with Fuzzy Sets," *Journal of Cybernetics*, vol. 3, pp. 58-73, 1973/01/01 1973.

[159] Y. Fukuyama and M. Sugeno, "A new method of choosing the number of clusters for the fuzzy c-means method," in *Proc. 5th Fuzzy Syst. Symp*, 1989.

[160] X. Xuanli Lisa and G. Beni, "A validity measure for fuzzy clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 13, pp. 841-847, 1991.

[161] The University of Sheffield. (2014, May 2014). *Sheffield WRGRID-ICEBERG*. Available: https://www.shef.ac.uk/wrgrid/iceberg

[162] S. Al-Sharhan, F. Karray, W. Gueaieb, and O. Basir, "Fuzzy entropy: A brief survey," Melbourne, 2001, pp. 1135-1139.

[163] R. Clausius, *The Mechanical Theory of Heat – with its Applications to the Steam Engine and to Physical Properties of Bodies*. London: John Van Voorst, 1867.

[164] J. C. Maxwell, *Theory of Heat*: Westport, Conn., Greenwood Press, 1871.

[165] C. E. Shannon, "A mathematical theory of communication," *SIGMOBILE Mob. Comput. Commun. Rev.*, vol. 5, pp. 3-55, 2001.

[166] A. De Luca and S. Termini, "A definition of a nonprobabilistic entropy in the setting of fuzzy sets theory," *Information and control*, vol. 20, pp. 301-312, 1972.

[167] R. Jensen and Q. Shen, "Fuzzy-rough data reduction with ant colony optimization," *Fuzzy Sets and Systems*, vol. 149, pp. 5-20, 2005.

[168] H. M. Lee, C. M. Chen, J. M. Chen, and Y. L. Jou, "An efficient fuzzy classifier with feature selection based on fuzzy entropy," *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics*, vol. 31, pp. 426-432, 2001.

[169] P. Yao, "Fuzzy rough set and information entropy based feature selection for credit scoring," Tianjin, 2009, pp. 247-251.

[170] K. Maertens, J. De Baerdemaeker, and R. Babuška, "Genetic polynomial regression as input selection algorithm for non-linear identification," *Soft Computing*, vol. 10, pp. 785-795, 2006.

[171] O. Gevaert, F. De Smet, D. Timmerman, Y. Moreau, and B. De Moor, "Predicting the prognosis of breast cancer by integrating clinical and microarray data with Bayesian networks," *BIOINFORMATICS*, vol. 22, pp. e184-e190, 2006.

[172] W. Kuo, T. Jenssen, A. Butte, L. Ohno-Machado, and I. Kohane, "Analysis of matched mRNA measurements from two different microarray technologies," *Bioinformatics*, vol. 18, pp. 405 - 412, 2002.

[173] S. Mitchell, K. Brown, M. Henry, M. Mintz, D. Catchpoole, B. LaFleur, and D. Stephan, "Inter-platform comparability of microarrays in acute lymphoblastic leukemia," *BMC Genomics*, vol. 5, p. 71, 2004.

[174] P. Warnat, R. Eils, and B. Brors, "Cross-platform analysis of cancer microarray data improves gene expression based classification of phenotypes," *BMC Bioinformatics*, vol. 6, p. 265, 2005.

[175] A. C. Lorena, I. G. Costa, N. Spolaôr, and M. C. P. De Souto, "Analysis of complexity indices for classification problems: Cancer gene expression data," *Neurocomputing*, vol. 75, pp. 33-42, 2012.

[176] H. Tin Kam and M. Basu, "Complexity measures of supervised classification problems," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 24, pp. 289-300, 2002.

[177] J. H. Phan, A. N. Young, and M. D. Wang, "Robust microarray meta-analysis identifies differentially expressed genes for clinical prediction," *The Scientific World Journal*, vol. 2012, 2012.

[178] J. K. Choi, J. Y. Choi, D. G. Kim, D. W. Choi, B. Y. Kim, K. H. Lee, Y. I. Yeom, H. S. Yoo, O. J. Yoo, and S. Kim, "Integrative analysis of multiple gene expression profiles applied to liver cancer study," *FEBS Letters*, vol. 565, pp. 93-100, 2004.

[179] D. R. Rhodes, T. R. Barrette, M. A. Rubin, D. Ghosh, and A. M. Chinnaiyan, "Meta-analysis of microarrays: interstudy validation of gene expression profiles reveals pathway dysregulation in prostate cancer," *Cancer Res*, vol. 62, pp. 4427-33, 2002.

[180] T. Park, S. G. Yi, Y. K. Shin, and S. Lee, "Combining multiple microarrays in the presence of controlling variables," *Bioinformatics*, vol. 22, pp. 1682-9, 2006.

[181] J. K. Choi, U. Yu, S. Kim, and O. J. Yoo, "Combining multiple microarray studies and modeling interstudy variation," *Bioinformatics*, vol. 19, pp. i84-90, 2003.

[182] Y. Tu, G. Stolovitzky, and U. Klein, "Quantitative noise analysis for gene expression microarray experiments," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, pp. 14031-14036, 2002.

[183] H. Yue, P. S. Eastman, B. B. Wang, J. Minor, M. H. Doctolero, R. L. Nuttall, R. Stack, J. W. Becker, J. R. Montgomery, M. Vainer, and R. Johnston, "An evaluation of the performance of cDNA microarrays for detecting changes in global mRNA expression," *Nucleic acids research*, vol. 29, pp. E41-41, 2001.

[184] V. E. Velculescu, L. Zhang, W. Zhou, J. Vogelstein, M. A. Basrai, D. E. Bassett Jr, P. Hieter, B. Vogelstein, and K. W. Kinzler, "Characterization of the yeast transcriptome," *Cell*, vol. 88, pp. 243-251, 1997.

[185] S. Blackshaw, W. P. Kuo, P. J. Park, M. Tsujikawa, J. M. Gunnersen, H. S. Scott, W. M. Boon, S. S. Tan, and C. L. Cepko, "MicroSAGE is highly representative and reproducible but reveals major differences in gene expression among samples obtained from similar tissues," *Genome biology,* vol. 4, 2003.

[186] P. Stafford and M. Brun, "Three methods for optimization of cross-laboratory and cross-platform microarray expression data," *Nucleic acids research,* vol. 35, 2007.

[187] C. Cheadle, M. P. Vawter, W. J. Freed, and K. G. Becker, "Analysis of microarray data using Z score transformation," *Journal of Molecular Diagnostics,* vol. 5, pp. 73-81, 2003.

[188] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub, "A molecular signature of metastasis in primary solid tumors," *Nature Genetics,* vol. 33, pp. 49-54, 2003.

[189] K. B. Hwang, S. W. Kong, S. A. Greenberg, and P. J. Park, "Combining gene expression data from different generations of oligonucleotide arrays," *BMC Bioinformatics,* vol. 5, 2004.

[190] G. Bloom, I. Yang, D. Boulware, K. Kwong, D. Coppola, S. Eschrich, J. Quackenbush, and T. Yeatman, "Multi-platform, multi-site, microarray-based human tumour classification," *Am J Pathol,* vol. 164, pp. 9 - 16, 2004.

[191] H. Liu, F. Hussain, C. Tan, and M. Dash, "Discretization: An enabling technique," *Data Mining and Knowledge Discovery*, vol. 6, pp. 393 - 423, 2002.

[192] C. K. Sarmah and S. Samarasinghe, "Microarray data integration: Frameworks and a list of underlying issues," *Current Bioinformatics,* vol. 5, pp. 280-289, 2010.

[193] A. Ramasamy, A. Mondry, C. C. Holmes, and D. G. Altman, "Key issues in conducting a meta-analysis of gene expression microarray datasets," *PLoS medicine,* vol. 5, 2008.

[194] W. McCulloch and W. Pitts, "A logical calculus of the ideas immanent in nervous activity," *The bulletin of mathematical biophysics,* vol. 5, pp. 115-133, 1943/12/01 1943.

[195] J. P. Florido, H. Pomares, I. Rojas, A. Guillén, F. M. Ortuno, and J. M. Urquiza, "An effective, practical and low computational cost framework for the integration of heterogeneous data to predict functional associations between proteins by means of artificial neural networks," *Neurocomputing,* vol. 121, pp. 64-78, 2013.

[196] M. E. Blazadonakis, M. E. Zervakis, and D. Kafetzopoulos, "Complementary gene signature integration in multiplatform microarray experiments," *IEEE Transactions on Information Technology in Biomedicine,* vol. 15, pp. 155-163, 2011.

Appendix A

Results with 1000 inputs and 5 rules

Ninety (90) of the included genes that are related with bladder cancer are: Gene NOS2, PRKCD, ALOX5, RBM3, CREBBP, RANBP3, GLB1, UQCRC2, CSE1L, GAST, HNRNPM, CSRP1, ASB8, SH3BGR, S100A7, MIF, HADHB, FER, HIF1AN, RGS2, LDHB, ENAH, MXD1, PRKACA, F3, ADA, CCL5, MVP, UPK2, IRF1, AR, TDG, CALU, MGMT, POLD1, AFF3, BCHE, XBP1, USP22, S100A10, ST3GAL1, VEGFA, HLA-B, CDCP1, GH2, ALDH1A2, CEACAM3, GALNS, HOPX, ADRB3, SERPINE1, STAT6, KIF20B, DNM2, FOXC1, SEMA3C, GABPB1, SOAT2, TNC, PPARG, TRADD, ITFG1, EGR1, NR3C1, OPCML, NAE1, ARHGDIB, POU4F2, MAP3K11, SEMA3A, JUN, MUC5B, TBX21, IGF1, DMBT1, DUSP10, ERCC3, IRF2, TP63, LGALS1, KLF4, PRDX3, DRAM1, MMP9, RECQL4, RPA2, LRP6, CAV2, AMACR, ME1.

Results with 2000 inputs and 5 rules

Additionally, 181 of the included genes that are related with bladder cancer are: Gene SUN2, SCYL2, BLNK, RHOC, HSPD1, FAAH, MMP11, TXNRD1, POLR2C, LZTS1, RNASE4, HLA-DRA, CDT1, AREG, IGFBP6, PLD3, IMP3, TLR4, GRHL2, NOTCH1, LEPREL4, L1TD1, PDE2A, ZNF135, MED12, CANT1, RNF43, HNF1A, EPCAM, FILIP1L, CALCOCO2, CCNE2, ATP5B, LGALS8, CST3, MSN, TEK, SCAMP3, ATPAF2, FIBP, PPFIBP1, SOCS3, TRPV2, CTH, CD59, EDNRA, S100A8, KAT7, ALPK3, LUM, MTHFR, USH2A, CYR61, XDH, PCDH17, ACVRL1, SLCO1B3, WWOX, RARB, SETD2, TERF2, ENPEP, NNMT, OXCT1,ERO1LB,

SULT1A1, YBX1, IL1R1,GDF9, TP73, FANCE, SHH, TP53I3, SULT1A2, HNF1B, DDX21, RCOR1, MAP4, RHOA, SELP, ACADM, EREG, MME, RAB15, ATPIF1, IL12A, GLIPR1, BMP6, CD44, SLIT2, FHIT, OS9, DIO2, BIRC7, FGF6, EIF3I, CD81, PSG1, HSPA4, SOX9, NID2, PLK2, HSP90B1, PBX3, RHOT1, FLNA, NQO1, CCL21, FGF1, MUC7, TERF1, OGT, NR5A1, TERT, ENO2, AKR1B10, TPM2, PSMB5, TAGLN, LY75, SRRM1, NCL, ADRA1A, TOP2A, F11R, ATP1A1, KIR2DS4, PDGFRB, OPRD1, GEMIN4, ESRP2, THBS2, DSP,TNKS, NFIL3, RANGAP1, PRKCSH, DES, ABCC4, 7-Sep, CAV1, COL1A1, STC1, HSPBAP1, WDR47, DFFB, TOMM34, CGA, TRAP1, TRIT1, POU5F1B, ZNF143, TAPBP, GLS, POLQ, GSTM5, STAM, CASK, CAT, FOXA1, FAH, PLAU, ACAT2, RACGAP1, GTF2I, MYH9, NFKBIB, PIN1, NR2F6, RGS6, GAA, CXADR, PAK6, RPA3, ADD3, IGFBP2, HRH1, CD36, MT3, VEGFC, CASP8.

Results with 5000 inputs and 5 rules

The Gene Signature obtained for the prediction of Survival contains the 5000 top ranked genes. A total of eight hundred and thirty eight (838) of the included genes that are related to bladder cancer are shown in Appendix A.

Table A.1: Genes related with Bladder Cancer

	Genes related with Bladder Cancer
1	2,4-dienoyl CoA reductase 1, mitochondrial
2	24-dehydrocholesterol reductase
3	5-hydroxytryptamine (serotonin) receptor 1B, G protein-coupled
4	5-hydroxytryptamine (serotonin) receptor 2A, G protein-coupled
5	A kinase (PRKA) anchor protein 13
6	actin related protein 2/3 complex, subunit 1B, 41kDa
7	actin related protein 2/3 complex, subunit 2, 34kDa
8	actin, alpha, cardiac muscle 1
9	actinin, alpha 1
10	activin A receptor type II-like 1
11	ADAM metallopeptidase domain 12
12	ADAM metallopeptidase with thrombospondin type 1 motif, 1
13	adducin 3 (gamma)

14	adenosine deaminase
15	adenylate cyclase 10 (soluble)
16	adrenergic, beta, receptor kinase 2
17	adrenoceptor alpha 1A
18	adrenoceptor beta 3
19	AE binding protein 1
20	AHNAK nucleoprotein
21	aldehyde dehydrogenase 1 family, member A2
22	aldehyde dehydrogenase 1 family, member A3
23	aldehyde dehydrogenase 3 family, member A2
24	aldo-keto reductase family 1, member A1 (aldehyde reductase)
25	aldo-keto reductase family 1, member B10 (aldose reductase)
26	aldolase A, fructose-bisphosphate
27	alkaline phosphatase, placental
28	alpha-1-microglobulin/bikunin precursor
29	alpha-methylacyl-CoA racemase
30	aminoacyl tRNA synthetase complex-interacting multifunctional protein 2
31	aminolevulinate, delta-, synthase 1
32	amphiregulin
33	amyloid beta (A4) precursor protein
34	anaphase promoting complex subunit 5
35	androgen receptor
36	angiogenin, ribonuclease, RNase A family, 5
37	angiopoietin-like 2
38	angiotensin II receptor, type 1
39	annexin A1
40	annexin A10
41	annexin A5
42	anthrax toxin receptor 1
43	apolipoprotein A-I
44	apolipoprotein B mRNA editing enzyme, catalytic polypeptide-like 3A
45	apoptotic chromatin condensation inducer 1
46	apoptotic peptidase activating factor 1
47	aquaporin 8
48	arachidonate 5-lipoxygenase
49	arginine and glutamate rich 1
50	argininosuccinate synthase 1
51	ARP2 actin-related protein 2 homolog (yeast)
52	ataxia telangiectasia mutated
53	ATP citrate lyase
54	ATP synthase, H+ transporting, mitochondrial F1 complex, beta polypeptide
55	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit B1
56	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit C2 (subunit 9)
57	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit d

58	ATP synthase, H+ transporting, mitochondrial Fo complex, subunit F6
59	ATPase inhibitory factor 1
60	ATPase, Ca++ transporting, cardiac muscle, slow twitch 2
61	ATPase, Cu++ transporting, alpha polypeptide
62	ATPase, Cu++ transporting, beta polypeptide
63	ATPase, Na+/K+ transporting, alpha 1 polypeptide
64	ATP-binding cassette, sub-family A (ABC1), member 7
65	ATP-binding cassette, sub-family B (MDR/TAP), member 6
66	ATP-binding cassette, sub-family C (CFTR/MRP), member 3
67	ATP-binding cassette, sub-family C (CFTR/MRP), member 4
68	aurora kinase A
69	AXL receptor tyrosine kinase
70	baculoviral IAP repeat containing 7
71	basic helix-loop-helix family, member e40
72	basic leucine zipper and W2 domains 1
73	B-cell CLL/lymphoma 11A (zinc finger protein)
74	BCL2-antagonist/killer 1
75	BCL2-associated agonist of cell death
76	BCL2-associated X protein
77	BCL2-like 1
78	BCL2-like 10 (apoptosis facilitator)
79	beta-1,3-glucuronyltransferase 1 (glucuronosyltransferase P)
80	betaine--homocysteine S-methyltransferase
81	Bloom syndrome, RecQ helicase-like
82	bone marrow stromal cell antigen 2
83	bone morphogenetic protein 6
84	bone morphogenetic protein 7
85	bone morphogenetic protein receptor, type II (serine/threonine kinase)
86	bradykinin receptor B2
87	brain-derived neurotrophic factor
88	BRCA1 associated RING domain 1
89	breakpoint cluster region
90	breast cancer 1, early onset
91	breast cancer metastasis suppressor 1
92	butyrylcholinesterase
93	Ca++-dependent secretion activator
94	cadherin 11, type 2, OB-cadherin (osteoblast)
95	cadherin 2, type 1, N-cadherin (neuronal)
96	cadherin 3, type 1, P-cadherin (placental)
97	calcium binding and coiled-coil domain 2
98	calcium/calmodulin-dependent serine protein kinase (MAGUK family)
99	calcium-sensing receptor
100	caldesmon 1
101	calmodulin binding transcription activator 1

102	calpain 1, (mu/I) large subunit
103	calpain 2, (m/II) large subunit
104	calpastatin
105	calumenin
106	cannabinoid receptor 1 (brain)
107	CAP, adenylate cyclase-associated protein 1 (yeast)
108	carboxypeptidase A3 (mast cell)
109	carboxypeptidase E
110	carcinoembryonic antigen-related cell adhesion molecule 3
111	carcinoembryonic antigen-related cell adhesion molecule 5
112	carcinoembryonic antigen-related cell adhesion molecule 6 (non-specific cross reacting antigen)
113	casein kinase 1, alpha 1
114	casein kinase 1, delta
115	caspase 10, apoptosis-related cysteine peptidase
116	caspase 2, apoptosis-related cysteine peptidase
117	caspase 8, apoptosis-related cysteine peptidase
118	catalase
119	catenin (cadherin-associated protein), beta 1, 88kDa
120	cathelicidin antimicrobial peptide
121	cathepsin H
122	cathepsin S
123	caudal type homeobox 1
124	caveolin 1, caveolae protein, 22kDa
125	caveolin 2
126	CCCTC-binding factor (zinc finger protein)
127	CD164 molecule, sialomucin
128	CD36 molecule (thrombospondin receptor)
129	CD3e molecule, epsilon associated protein
130	CD4 molecule
131	CD40 molecule, TNF receptor superfamily member 5
132	CD44 molecule (Indian blood group)
133	CD59 molecule, complement regulatory protein
134	CD81 molecule
135	CD82 molecule
136	CD99 molecule
137	cellular retinoic acid binding protein 1
138	chaperonin containing TCP1, subunit 2 (beta)
139	chaperonin containing TCP1, subunit 3 (gamma)
140	chaperonin containing TCP1, subunit 4 (delta)
141	chaperonin containing TCP1, subunit 6A (zeta 1)
142	chaperonin containing TCP1, subunit 7 (eta)
143	checkpoint kinase 1
144	chemokine (C-C motif) ligand 2

145	chemokine (C-C motif) ligand 21
146	chemokine (C-C motif) ligand 22
147	chemokine (C-C motif) ligand 5
148	chemokine (C-X-C motif) ligand 1 (melanoma growth stimulating activity, alpha)
149	chitinase 3-like 1 (cartilage glycoprotein-39)
150	chloride channel accessory 2
151	chloride intracellular channel 4
152	cholecystokinin B receptor
153	choline kinase alpha
154	choline O-acetyltransferase
155	cholinergic receptor, muscarinic 3
156	chromodomain helicase DNA binding protein 1-like
157	citrate synthase
158	clathrin, heavy chain (Hc)
159	clathrin, light chain A
160	claudin 10
161	CNDP dipeptidase 2 (metallopeptidase M20 family)
162	coagulation factor II (thrombin) receptor-like 2
163	coagulation factor III (thromboplastin, tissue factor)
164	coatomer protein complex, subunit alpha
165	collagen, type I, alpha 1
166	collagen, type V, alpha 1
167	collagen, type VI, alpha 1
168	collagen, type VI, alpha 2
169	collagen, type VI, alpha 3
170	collagen, type VII, alpha 1
171	collagen, type XVI, alpha 1
172	colony stimulating factor 1 (macrophage)
173	complement component 1, q subcomponent binding protein
174	contactin 2 (axonal)
175	cortactin
176	coxsackie virus and adenovirus receptor
177	c-ros oncogene 1 , receptor tyrosine kinase
178	crystallin, alpha B
179	CSE1 chromosome segregation 1-like (yeast)
180	C-terminal binding protein 1
181	cyclin A1
182	cyclin A2
183	cyclin B2
184	cyclin E2
185	cyclin G1
186	cyclin L1
187	cyclin-dependent kinase 2 associated protein 1
188	cyclin-dependent kinase 4

189	cyclin-dependent kinase 7
190	cyclin-dependent kinase inhibitor 1C (p57, Kip2)
191	cystathionase (cystathionine gamma-lyase)
192	cystatin C
193	cysteine and glycine-rich protein 1
194	cysteine-rich, angiogenic inducer, 61
195	cystic fibrosis transmembrane conductance regulator (ATP-binding cassette sub-family C, member 7)
196	cytochrome c oxidase subunit VIc
197	cytochrome P450, family 1, subfamily A, polypeptide 2
198	cytochrome P450, family 2, subfamily A, polypeptide 13
199	cytochrome P450, family 2, subfamily C, polypeptide 19
200	cytochrome P450, family 2, subfamily D, polypeptide 6
201	cytochrome P450, family 4, subfamily B, polypeptide 1
202	cytochrome P450, family 7, subfamily B, polypeptide 1
203	damage-specific DNA binding protein 2, 48kDa
204	dCMP deaminase
205	DEAH (Asp-Glu-Ala-His) box polypeptide 35
206	death associated protein 3
207	death-associated protein
208	death-associated protein kinase 1
209	death-associated protein kinase 2
210	deiodinase, iodothyronine, type I
211	deleted in malignant brain tumors 1
212	desmin
213	desmoglein 3
214	desmoplakin
215	destrin (actin depolymerizing factor)
216	dihydropyrimidinase-like 3
217	DNA (cytosine-5-)-methyltransferase 1
218	DNA fragmentation factor, 40kDa, beta polypeptide (caspase-activated DNase)
219	DnaJ (Hsp40) homolog, subfamily B, member 1
220	dopamine beta-hydroxylase (dopamine beta-monooxygenase)
221	dopamine receptor D2
222	drosha, ribonuclease type III
223	dual specificity phosphatase 1
224	dynamin 2
225	E2F transcription factor 1
226	E2F transcription factor 3
227	early growth response 1
228	ectonucleoside triphosphate diphosphohydrolase 3
229	ectonucleoside triphosphate diphosphohydrolase 4
230	EGF containing fibulin-like extracellular matrix protein 1
231	EGF containing fibulin-like extracellular matrix protein 2

232	EGF-like repeats and discoidin I-like domains 3
233	EGF-like-domain, multiple 6
234	ELK1, member of ETS oncogene family
235	enabled homolog (Drosophila)
236	endothelin 1
237	endothelin receptor type A
238	endothelin receptor type B
239	enolase 2 (gamma, neuronal)
240	EPH receptor B4
241	ephrin-A1
242	epidermal growth factor receptor
243	epidermal growth factor receptor pathway substrate 8
244	epiregulin
245	epithelial cell adhesion molecule
246	epithelial membrane protein 1
247	epoxide hydrolase 2, cytoplasmic
248	ERGIC and golgi 3
249	estrogen receptor 1
250	estrogen receptor 2 (ER beta)
251	eukaryotic translation elongation factor 2
252	eukaryotic translation initiation factor 3, subunit I
253	eukaryotic translation initiation factor 4 gamma, 2
254	eukaryotic translation initiation factor 4E binding protein 1
255	eukaryotic translation initiation factor 4H
256	eukaryotic translation initiation factor 5
257	eukaryotic translation initiation factor 5A
258	ezrin
259	family with sequence similarity 215, member A (non-protein coding)
260	Fanconi anemia, complementation group A
261	Fanconi anemia, complementation group F
262	far upstream element (FUSE) binding protein 1
263	far upstream element (FUSE) binding protein 3
264	fatty acid amide hydrolase
265	fatty acid binding protein 1, liver
266	FBJ murine osteosarcoma viral oncogene homolog B
267	fer (fps/fes related) tyrosine kinase
268	ferrochelatase
269	fibrillarin
270	fibroblast growth factor 1 (acidic)
271	fibroblast growth factor 6
272	fibroblast growth factor 7
273	fibroblast growth factor receptor 3
274	filamin A interacting protein 1-like
275	filamin A, alpha

276	FK506 binding protein 4, 59kDa
277	folate hydrolase (prostate-specific membrane antigen) 1
278	folate receptor 1 (adult)
279	follistatin-like 1
280	forkhead box A1
281	forkhead box C1
282	fragile histidine triad
283	fructose-1,6-bisphosphatase 1
284	fucosidase, alpha-L- 1, tissue
285	fucosyltransferase 3 (galactoside 3(4)-L-fucosyltransferase, Lewis blood group)
286	fucosyltransferase 6 (alpha (1,3) fucosyltransferase)
287	fused in sarcoma
288	FXYD domain containing ion transport regulator 1
289	FXYD domain containing ion transport regulator 3
290	FYN oncogene related to SRC, FGR, YES
291	G protein-coupled estrogen receptor 1
292	galactosamine (N-acetyl)-6-sulfate sulfatase
293	gamma-glutamyl hydrolase (conjugase, folylpolygammaglutamyl hydrolase)
294	gap junction protein, alpha 1, 43kDa
295	gastrin
296	GDP-mannose 4,6-dehydratase
297	gem (nuclear organelle) associated protein 4
298	general transcription factor IIi
299	GLI family zinc finger 1
300	GLI family zinc finger 3
301	GLI pathogenesis-related 1
302	glioma tumor suppressor candidate region gene 2
303	glucan (1,4-alpha-), branching enzyme 1
304	glutamic-pyruvate transaminase (alanine aminotransferase)
305	glutaminase
306	glutamyl aminopeptidase (aminopeptidase A)
307	glutamyl-prolyl-tRNA synthetase
308	glutathione peroxidase 1
309	glutathione peroxidase 2 (gastrointestinal)
310	glutathione peroxidase 3 (plasma)
311	glutathione reductase
312	glutathione S-transferase alpha 3
313	glutathione S-transferase alpha 4
314	glutathione S-transferase pi 1
315	glyceraldehyde-3-phosphate dehydrogenase
316	glycoprotein (transmembrane) nmb
317	glycoprotein hormones, alpha polypeptide
318	golgi phosphoprotein 3 (coat-protein)
319	gonadotropin-releasing hormone receptor

320	granulin
321	growth arrest and DNA-damage-inducible, beta
322	growth differentiation factor 15
323	growth differentiation factor 9
324	growth hormone 2
325	growth hormone inducible transmembrane protein
326	growth hormone receptor
327	guanine nucleotide binding protein (G protein), gamma 11
328	guanosine monophosphate reductase 2
329	guanylate kinase 1
330	H2A histone family, member X
331	H3 histone, family 3A
332	heat shock 60kDa protein 1 (chaperonin)
333	heat shock 70kDa protein 4
334	heat shock protein 90kDa alpha (cytosolic), class A member 1
335	heat shock protein 90kDa alpha (cytosolic), class B member 1
336	heat shock protein 90kDa beta (Grp94), member 1
337	hemochromatosis
338	heparanase
339	heparin-binding EGF-like growth factor
340	hepatocyte growth factor (hepapoietin A; scatter factor)
341	heterogeneous nuclear ribonucleoprotein D (AU-rich element RNA binding protein 1, 37kDa)
342	heterogeneous nuclear ribonucleoprotein D-like
343	heterogeneous nuclear ribonucleoprotein F
344	heterogeneous nuclear ribonucleoprotein H1 (H)
345	high density lipoprotein binding protein
346	high mobility group box 1
347	histamine receptor H1
348	histone deacetylase 9
349	HNF1 homeobox B
350	HOP homeobox
351	HSPB (heat shock 27kDa) associated protein 1
352	HtrA serine peptidase 1
353	hyaluronan synthase 2
354	hyaluronan-mediated motility receptor (RHAMM)
355	hyaluronoglucosaminidase 1
356	hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase (trifunctional protein), beta subunit
357	hydroxyprostaglandin dehydrogenase 15-(NAD)
358	hydroxysteroid (17-beta) dehydrogenase 2
359	hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)
360	hypoxia inducible factor 1, alpha subunit inhibitor
361	immunoglobulin superfamily containing leucine-rich repeat
362	InaD-like (Drosophila)

363	inhibitor of DNA binding 2, dominant negative helix-loop-helix protein
364	inhibitor of DNA binding 3, dominant negative helix-loop-helix protein
365	inositol 1,4,5-trisphosphate receptor, type 1
366	inositol hexakisphosphate kinase 2
367	insulin receptor
368	insulin receptor substrate 1
369	insulin-like growth factor 1 (somatomedin C)
370	insulin-like growth factor 1 receptor
371	insulin-like growth factor 2 mRNA binding protein 3
372	insulin-like growth factor binding protein 2, 36kDa
373	insulin-like growth factor binding protein 3
374	insulin-like growth factor binding protein 6
375	integrin alpha FG-GAP repeat containing 1
376	integrin, alpha 2 (CD49B, alpha 2 subunit of VLA-2 receptor)
377	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)
378	integrin, alpha L (antigen CD11A (p180), lymphocyte function-associated antigen 1; alpha polypeptide)
379	integrin, alpha V
380	integrin, beta 5
381	integrin-linked kinase
382	intercellular adhesion molecule 2
383	interferon regulatory factor 1
384	interferon regulatory factor 3
385	interferon, gamma
386	interleukin 1 receptor antagonist
387	interleukin 1 receptor, type I
388	interleukin 1, alpha
389	interleukin 12A (natural killer cell stimulatory factor 1, cytotoxic lymphocyte maturation factor 1, p35)
390	interleukin 13
391	interleukin 17 receptor A
392	interleukin 17A
393	interleukin 5 (colony-stimulating factor, eosinophil)
394	interleukin 6 (interferon, beta 2)
395	interleukin 6 signal transducer (gp130, oncostatin M receptor)
396	interleukin 7 receptor
397	interleukin 8
398	interleukin enhancer binding factor 3, 90kDa
399	ISL LIM homeobox 1
400	Janus kinase 1
401	jumping translocation breakpoint
402	jun proto-oncogene
403	karyopherin (importin) beta 1
404	keratin 1
405	keratin 14

406	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 4
407	kinesin family member 16B
408	kinesin family member 20B
409	Kruppel-like factor 4 (gut)
410	Kruppel-like factor 5 (intestinal)
411	lactate dehydrogenase B
412	laminin, alpha 3
413	laminin, beta 2 (laminin S)
414	laminin, gamma 2
415	laminin, gamma 3
416	latent transforming growth factor beta binding protein 1
417	lecithin retinol acyltransferase (phosphatidylcholine--retinol O-acyltransferase)
418	lectin, galactoside-binding, soluble, 1
419	lectin, galactoside-binding, soluble, 3 binding protein
420	lectin, galactoside-binding, soluble, 8
421	lectin, galactoside-binding, soluble, 9
422	legumain
423	leucine zipper, putative tumor suppressor 1
424	ligase IV, DNA, ATP-dependent
425	LIM domain and actin binding 1
426	LIM domain only 2 (rhombotin-like 1)
427	LINE-1 type transposase domain containing 1
428	lipase, hepatic
429	lipocalin 2
430	lipopolysaccharide-induced TNF factor
431	low density lipoprotein receptor-related protein 5
432	low density lipoprotein receptor-related protein 6
433	low density lipoprotein receptor-related protein associated protein 1
434	lumican
435	lymphocyte antigen 6 complex, locus E
436	lymphocyte antigen 75
437	lysosomal-associated membrane protein 2
438	lysozyme
439	lysyl-tRNA synthetase
440	macrophage migration inhibitory factor (glycosylation-inhibiting factor)
441	major histocompatibility complex, class I, A
442	major histocompatibility complex, class I, B
443	major histocompatibility complex, class II, DR alpha
444	major intrinsic protein of lens fiber
445	major vault protein
446	mannose-binding lectin (protein C) 2, soluble
447	mannose-P-dolichol utilization defect 1
448	MAP7 domain containing 1
449	MAP-kinase activating death domain

450	matrix Gla protein
451	matrix metallopeptidase 10 (stromelysin 2)
452	matrix metallopeptidase 11 (stromelysin 3)
453	matrix metallopeptidase 2 (gelatinase A, 72kDa gelatinase, 72kDa type IV collagenase)
454	matrix metallopeptidase 9 (gelatinase B, 92kDa gelatinase, 92kDa type IV collagenase)
455	mechanistic target of rapamycin (serine/threonine kinase)
456	melanoma antigen family D, 2
457	membrane metallo-endopeptidase
458	metallothionein 1F
459	metallothionein 1X
460	metallothionein 2A
461	metallothionein 3
462	metastasis suppressor 1
463	methyl CpG binding protein 2 (Rett syndrome)
464	methylenetetrahydrofolate reductase (NAD(P)H)
465	microtubule-associated protein 4
466	midkine (neurite growth-promoting factor 2)
467	mitochondrial calcium uptake 1
468	mitogen-activated protein kinase 3
469	mitogen-activated protein kinase 8
470	mitogen-activated protein kinase associated protein 1
471	mitogen-activated protein kinase kinase 2
472	mitogen-activated protein kinase kinase 4
473	mitogen-activated protein kinase kinase 7
474	moesin
475	mortality factor 4 like 2
476	motilin
477	mucin 16, cell surface associated
478	mucin 3A, cell surface associated
479	mucin 5B, oligomeric mucus/gel-forming
480	mucin 7, secreted
481	SDA1 domain containing 1
482	RAB5B, member RAS oncogene family
483	aldo-keto reductase family 1, member A1 (aldehyde reductase)
484	growth arrest-specific 7
485	downstream neighbor of SON
486	insulin induced gene 2
487	makorin ring finger protein 2
488	mitogen-activated protein kinase kinase kinase kinase 4
489	bystin-like
490	SPC25, NDC80 kinetochore complex component, homolog (S. cerevisiae)
491	small nucleolar RNA, H/ACA box 5B /// transforming growth factor beta regulator 4
492	transmembrane and ubiquitin-like domain containing 2
493	NKF3 kinase family member

494	gon-4-like (C. elegans)
495	transmembrane protease, serine 11D
496	BAH domain and coiled-coil containing 1
497	transient receptor potential cation channel, subfamily C, member 1
498	ArfGAP with SH3 domain, ankyrin repeat and PH domain 2
499	peptidase domain containing associated with muscle regeneration 1
500	myosin, light chain 12A, regulatory, non-sarcomeric
501	A kinase (PRKA) anchor protein 17A
502	transmembrane protein 120B
503	HLA complex group 26 (non-protein coding)
504	nitric oxide synthase 2, inducible
505	uncharacterized FLJ13197
506	proteasome (prosome, macropain) subunit, beta type, 10
507	protein kinase C, delta
508	KN motif and ankyrin repeat domains 2
509	transmembrane protein 126B
510	CTD (carboxy-terminal domain, RNA polymerase II, polypeptide A) phosphatase, subunit 1
511	paired related homeobox 2
512	thiosulfate sulfurtransferase (rhodanese)
513	visinin-like 1
514	arachidonate 5-lipoxygenase
515	collagen, type VI, alpha 3
516	DEAD (Asp-Glu-Ala-Asp) box polypeptide 19A
517	calcineurin binding protein 1
518	dickkopf 3 homolog (Xenopus laevis)
519	GTP binding protein 3 (mitochondrial)
520	RNA binding motif (RNP1, RRM) protein 3
521	DnaJ (Hsp40) homolog, subfamily C, member 28
522	ABHD14A-ACY1 readthrough (non-protein coding) /// aminoacylase 1
523	CREB binding protein
524	chromosome 5 open reading frame 25 pseudogene
525	lysine-rich coiled-coil 1
526	ubiquitin-like 4A
527	LIM homeobox 3
528	HAUS augmin-like complex, subunit 4 /// microRNA 4707
529	complement component 8, alpha polypeptide
530	zinc finger protein 329
531	integrin, beta 5
532	Yes-associated protein 1
533	neuronal pentraxin I
534	FRY-like
535	fermitin family member 1
536	SMAD specific E3 ubiquitin protein ligase 2

537	solute carrier family 7 (orphan transporter), member 4
538	uncharacterized LOC100508797
539	chromosome 3 open reading frame 36
540	myomesin 1, 185kDa
541	small proline-rich protein 1B
542	MOK protein kinase
543	tumor necrosis factor receptor superfamily, member 10d, decoy with truncated death domain
544	ring finger protein 114
545	spermine synthase
546	N-acetylated alpha-linked acidic dipeptidase-like 1
547	secretoglobin, family 2A, member 1
548	RAN binding protein 3
549	MANSC domain containing 1
550	myosin, light chain 1, alkali; skeletal, fast
551	zinc finger protein 711
552	RALY RNA binding protein-like
553	adrenoceptor alpha 2A
554	RAB8A, member RAS oncogene family
555	chromodomain helicase DNA binding protein 8
556	potassium channel tetramerisation domain containing 15
557	four and a half LIM domains 5
558	galactosidase, beta 1
559	ubiquinol-cytochrome c reductase core protein II
560	thromboxane A2 receptor
561	transmembrane protein 30A
562	TRM1 tRNA methyltransferase 1 homolog (S. cerevisiae)
563	leucine rich repeat containing 41
564	CSE1 chromosome segregation 1-like (yeast)
565	protein phosphatase 1, regulatory subunit 16B
566	early growth response 3
567	complexin 2
568	phosphatidylinositol-3,4,5-trisphosphate-dependent Rac exchange factor 2
569	ATP-binding cassette, sub-family B (MDR/TAP), member 8
570	zinc finger CCCH-type containing 3
571	methyl-CpG binding domain protein 1
572	RNA binding motif protein 14
573	pleckstrin homology domain containing, family G (with RhoGef domain) member 6
574	secretoglobin, family 1A, member 1 (uteroglobin)
575	TM2 domain containing 3
576	kallikrein-related peptidase 13
577	ankyrin repeat and SOCS box containing 6
578	coiled-coil domain containing 22
579	proteasome (prosome, macropain) inhibitor subunit 1 (PI31)

580	immunoglobulin lambda light chain-like
581	single-stranded DNA binding protein 2
582	gastrin
583	cyclin-dependent kinase 10
584	homeobox A6
585	protein phosphatase 1, regulatory subunit 9A
586	heterogeneous nuclear ribonucleoprotein M
587	astrotactin 2
588	tumor necrosis factor receptor superfamily, member 12A
589	tyrosyl-tRNA synthetase 2, mitochondrial
590	zinc finger, CW type with PWWP domain 1
591	cysteine and glycine-rich protein 1
592	nascent polypeptide-associated complex alpha subunit 2
593	myosin light chain kinase
594	DiGeorge syndrome critical region gene 11 (non-protein coding)
595	glutamate receptor, ionotropic, N-methyl D-aspartate 2D
596	THAP domain containing 7
597	ankyrin repeat and SOCS box containing 8
598	DDB1 and CUL4 associated factor 10
599	chromodomain helicase DNA binding protein 1-like
600	centrosomal protein 250kDa
601	transcription elongation factor A (SII), 1
602	RWD domain containing 2A
603	F-box protein 5
604	tubulin, alpha 1c
605	MAP7 domain containing 3
606	transcription factor 4
607	chromosome 15 open reading frame 39
608	nucleolar protein 10
609	SH3 domain binding glutamic acid-rich protein
610	S100 calcium binding protein A7
611	fructosamine 3 kinase
612	protein phosphatase 3, catalytic subunit, alpha isozyme
613	leucine rich repeat containing 19
614	fibronectin leucine rich transmembrane protein 1
615	ribosomal protein L9
616	ADP-ribosylation factor-like 4A
617	MON1 homolog B (yeast)
618	cryptochrome 2 (photolyase-like)
619	exocyst complex component 7
620	ring finger protein 139
621	chromosome 19 open reading frame 80
622	adaptor-related protein complex 3, sigma 2 subunit /// C15orf38-AP3S2 readthrough
623	twinfilin, actin-binding protein, homolog 2 (Drosophila)

624	PBX/knotted 1 homeobox 1
625	macrophage migration inhibitory factor (glycosylation-inhibiting factor)
626	deoxyribonuclease I-like 2
627	frizzled family receptor 2
628	lysophosphatidic acid receptor 2
629	heterogeneous nuclear ribonucleoprotein H2 (H) /// RPL36A-HNRNPH2 readthrough
630	fucosyltransferase 5 (alpha (1,3) fucosyltransferase)
631	gonadotropin-releasing hormone 2
632	exosome component 1
633	gremlin 2
634	phosphorylase kinase, gamma 1 (muscle)
635	blocked early in transport 1 homolog (S. cerevisiae)-like
636	SMAD family member 7
637	T-cell lymphoma invasion and metastasis 2
638	hydroxyacyl-CoA dehydrogenase/3-ketoacyl-CoA thiolase/enoyl-CoA hydratase (trifunctional protein), beta subunit
639	fer (fps/fes related) tyrosine kinase
640	hypoxia inducible factor 1, alpha subunit inhibitor
641	regulator of G-protein signaling 2, 24kDa
642	lactate dehydrogenase B
643	vav 1 guanine nucleotide exchange factor
644	FtsJ methyltransferase domain containing 2
645	microRNA 1292 /// NOP56 ribonucleoprotein homolog (yeast) /// small nucleolar RNA, C/D box 110 /// small nucleolar RNA, C/D box 57 /// small nucleolar RNA, C/D box 86
646	enabled homolog (Drosophila)
647	MAX dimerization protein 1
648	SET binding protein 1
649	PR domain containing 4
650	partner of NOB1 homolog (S. cerevisiae)
651	G protein-coupled receptor 161
652	DEAD (Asp-Glu-Ala-Asp) box polypeptide 43
653	TBC1 domain family, member 29
654	killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 1 /// killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 2 /// killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 3 /// killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 4 /// killer cell immunoglobulin-like receptor, two domains, short cytoplasmic tail, 5 /// killer cell immunoglobulin-like receptor three domains long cytoplasmic tail 3
655	T-box 4
656	X antigen family, member 1A /// X antigen family, member 1B /// X antigen family, member 1C /// X antigen family, member 1D /// X antigen family, member 1E
657	polymerase (RNA) III (DNA directed) polypeptide E (80kD)
658	trafficking protein particle complex 12
659	ATP-binding cassette, sub-family C (CFTR/MRP), member 5
660	protein kinase, cAMP-dependent, catalytic, alpha
661	coagulation factor III (thromboplastin, tissue factor)

662	multimerin 2
663	N-acetylglucosamine-1-phosphate transferase, alpha and beta subunits
664	bicaudal C homolog 1 (Drosophila)
665	potassium channel, subfamily K, member 15
666	olfactory receptor, family 1, subfamily G, member 1
667	sphingomyelin phosphodiesterase 2, neutral membrane (neutral sphingomyelinase)
668	PH domain and leucine rich repeat protein phosphatase 2
669	midkine (neurite growth-promoting factor 2)
670	protein tyrosine phosphatase, non-receptor type 6
671	gasdermin D
672	pleckstrin homology-like domain, family A, member 2
673	zinc finger protein 292
674	adenosine deaminase
675	ATP-binding cassette, sub-family F (GCN20), member 1
676	misato homolog 1 (Drosophila) /// misato homolog 2 pseudogene
677	cytochrome c oxidase subunit VIIa polypeptide 1 (muscle)
678	myotubularin related protein 2
679	ER degradation enhancer, mannosidase alpha-like 3
680	amyloid beta (A4) precursor-like protein 1
681	chemokine (C-C motif) ligand 5
682	DEAD (Asp-Glu-Ala-Asp) box polypeptide 19A /// DEAD (Asp-Glu-Ala-Asp) box polypeptide 19B
683	spermatogenesis associated 1
684	major vault protein
685	5-hydroxytryptamine (serotonin) receptor 6, G protein-coupled
686	poly(rC) binding protein 1
687	chondroitin polymerizing factor
688	aspartyl aminopeptidase
689	enhancer of rudimentary homolog (Drosophila)
690	bolA homolog 1 (E. coli)
691	coiled-coil domain containing 144A
692	ARP1 actin-related protein 1 homolog B, centractin beta (yeast)
693	uroplakin 2
694	interferon regulatory factor 1
695	CD96 molecule
696	ubiquitin protein ligase E3 component n-recognin 5
697	androgen receptor
698	thymine-DNA glycosylase
699	endothelial PAS domain protein 1 /// uncharacterized LOC100652809
700	CDKN2A interacting protein
701	solute carrier family 10 (sodium/bile acid cotransporter family), member 2
702	glycerol-3-phosphate dehydrogenase 1-like
703	calumenin
704	keratin 12

705	Ras suppressor protein 1
706	recombination activating gene 2
707	G protein-coupled receptor 15
708	uncharacterized LOC257152
709	FtsJ RNA methyltransferase homolog 2 (E. coli)
710	origin recognition complex, subunit 3
711	O-6-methylguanine-DNA methyltransferase
712	polymerase (DNA directed), delta 1, catalytic subunit
713	regulatory factor X-associated ankyrin-containing protein
714	guanine nucleotide binding protein (G protein), alpha z polypeptide
715	CCR4-NOT transcription complex, subunit 2
716	hCG1732469
717	microfibrillar-associated protein 2
718	SND1 intronic transcript 1 (non-protein coding)
719	D site of albumin promoter (albumin D-box) binding protein
720	AF4/FMR2 family, member 3
721	SPO11 meiotic protein covalently bound to DSB homolog (S. cerevisiae)
722	akirin 1
723	ASMTL antisense RNA 1 (non-protein coding)
724	chaperonin containing TCP1, subunit 4 (delta)
725	guanine nucleotide binding protein (G protein), beta 5
726	methyltransferase like 9
727	uncharacterized LOC100506190
728	butyrylcholinesterase
729	slowmo homolog 2 (Drosophila)
730	transmembrane 4 L six family member 5
731	X-box binding protein 1
732	ubiquitin specific peptidase 22
733	dapper, antagonist of beta-catenin, homolog 1 (Xenopus laevis)
734	phosphatidylinositol transfer protein, membrane-associated 1
735	cell division cycle 123 homolog (S. cerevisiae)
736	olfactory receptor, family 1, subfamily F, member 1
737	S100 calcium binding protein A10
738	ST3 beta-galactoside alpha-2,3-sialyltransferase 1
739	nitrogen permease regulator-like 2 (S. cerevisiae)
740	lymphocyte antigen 6 complex, locus G6E (pseudogene)
741	IKAROS family zinc finger 5 (Pegasus)
742	purinergic receptor P2X, ligand-gated ion channel, 6
743	exocyst complex component 6B
744	polymerase (RNA) I polypeptide C, 30kDa
745	vascular endothelial growth factor A
746	major histocompatibility complex, class I, B
747	centromere protein F, 350/400kDa (mitosin)
748	SH3 domain containing, Ysc84-like 1 (S. cerevisiae)

749	CUB domain containing protein 1
750	proline synthetase co-transcribed homolog (bacterial)
751	growth hormone 2
752	epidermal growth factor receptor pathway substrate 8
753	acylphosphatase 2, muscle type
754	THO complex 2
755	aldehyde dehydrogenase 1 family, member A2
756	tripartite motif containing 8
757	carcinoembryonic antigen-related cell adhesion molecule 3
758	WD repeat domain 77
759	centrosomal protein 85kDa
760	aminopeptidase-like 1 /// STX16-NPEPL1 readthrough (non-protein coding)
761	zinc finger, DHHC-type containing 7
762	protein phosphatase, Mg2+/Mn2+ dependent, 1E
763	TAR (HIV-1) RNA binding protein 2
764	galactosamine (N-acetyl)-6-sulfate sulfatase
765	nucleoporin 98kDa
766	Rho GTPase activating protein 15
767	vaccinia related kinase 1
768	zinc finger, BED-type containing 4
769	thyrotropin-releasing hormone degrading enzyme
770	membrane-spanning 4-domains, subfamily A, member 4A
771	HOP homeobox
772	ATPase, Ca++ transporting, plasma membrane 1
773	adrenoceptor beta 3
774	smg-5 homolog, nonsense mediated mRNA decay factor (C. elegans)
775	vacuolar protein sorting 13 homolog D (S. cerevisiae)
776	SLIT and NTRK-like family, member 3
777	glucose 6 phosphatase, catalytic, 3
778	succinate-CoA ligase, GDP-forming, beta subunit
779	snail homolog 1 (Drosophila)
780	vacuolar protein sorting 33 homolog B (yeast)
781	serpin peptidase inhibitor, clade E (nexin, plasminogen activator inhibitor type 1), member 1
782	signal transducer and activator of transcription 6, interleukin-4 induced
783	RUN and FYVE domain containing 3
784	solute carrier family 6 (neurotransmitter transporter, glycine), member 5
785	signal sequence receptor, gamma (translocon-associated protein gamma)
786	uncoupling protein 2 (mitochondrial, proton carrier)
787	zinc finger protein 674
788	HIG1 hypoxia inducible domain family, member 1A
789	olfactory receptor, family 7, subfamily C, member 1
790	ankyrin repeat domain 34C
791	general transcription factor IIH, polypeptide 2B (pseudogene)

792	ATP5S-like
793	coiled-coil domain containing 81
794	hepatoma-derived growth factor, related protein 3
795	transmembrane channel-like 7
796	spastic paraplegia 11 (autosomal recessive)
797	kinesin family member 20B
798	190 kDa guanine nucleotide exchange factor
799	chromosome 11 open reading frame 24
800	dynamin 2
801	CREB/ATF bZIP transcription factor
802	mannosidase, alpha, class 2A, member 1
803	G patch domain containing 3
804	troponin I type 2 (skeletal, fast)
805	EPH receptor A3
806	CD3g molecule, gamma (CD3-TCR complex)
807	ring finger protein 32
808	zinc finger, MYND-type containing 8
809	forkhead box C1
810	phosphoribosyl pyrophosphate synthetase-associated protein 1
811	ATPase, aminophospholipid transporter, class I, type 8B, member 3
812	melanocortin 5 receptor
813	family with sequence similarity 168, member A
814	RAB21, member RAS oncogene family
815	guanylate cyclase 1, soluble, beta 3
816	sema domain, immunoglobulin domain (Ig), short basic domain, secreted, (semaphorin) 3C
817	cerebellar degeneration-related protein 2, 62kDa
818	NAG18 mRNA
819	GA binding protein transcription factor, beta subunit 1
820	late endosomal/lysosomal adaptor, MAPK and MTOR activator 2
821	Ral GTPase activating protein, beta subunit (non-catalytic)
822	Sin3A-associated protein, 130kDa
823	alkylglycerone phosphate synthase
824	collagen, type XIII, alpha 1
825	WD repeat domain 46
826	pleckstrin and Sec7 domain containing
827	sterol O-acyltransferase 2
828	tenascin C
829	crystallin, mu
830	pseudouridylate synthase 3
831	chromosome 1 open reading frame 50
832	lin-7 homolog B (C. elegans)
833	G protein-coupled receptor 98
834	peroxisome proliferator-activated receptor gamma

835	TNFRSF1A-associated via death domain
836	NUAK family, SNF1-like kinase, 1
837	pyruvate dehyrogenase phosphatase catalytic subunit 1
838	macrophage erythroblast attacher

Appendix B

Table B.1: Top Genes 25 Blaveri Stage

Rank	Gene Title
1	KIAA0914 gene product
2	fibromodulin
3	extracellular matrix protein 1
4	matrix metalloproteinase 11 (stromelysin 3)
5	MHC class II transactivator
6	EGF-containing fibulin-like extracellular matrix protein 1
7	integrin, alpha V (vitronectin receptor, alpha polypeptide, antigen CD51)
8	Microfibril-associated glycoprotein-2
9	S100 calcium-binding protein A7 (psoriasin 1)
10	aquaporin 5
11	membrane cofactor protein (CD46, trophoblast-lymphocyte cross-reactive antigen)
12	KIAA0494 gene product
13	AE-binding protein 1
14	natural killer cell group 7 sequence
15	inhibin, beta A (activin A, activin AB alpha polypeptide)
16	contactin associated protein 1
17	ESTs, Moderately similar to JC4969 pig-c protein [H.sapiens]
18	duodenal cytochrome b
19	KIAA1077 protein
20	Human clone 23719 mRNA sequence
21	myosin regulatory light chain interacting protein
22	retinoic acid receptor responder (tazarotene induced) 2
23	RAB31, member RAS oncogene family
24	endothelin receptor type A
25	guanine nucleotide binding protein (G protein), alpha 11 (Gq class)

Table B.2: Top Genes 25 Blaveri Grade

Rank	Gene Title
1	keratin 19
2	v-yes-1 Yamaguchi sarcoma viral related oncogene homolog
3	nuclear cap binding protein subunit 1, 80kD
4	major histocompatibility complex, class II, DR alpha

5	N-acetylneuraminic acid phosphate synthase
6	erythrocyte membrane protein band 7.2 (stomatin)
7	regulator of G-protein signalling 10
8	FXYD domain containing ion transport regulator 3
9	pre-alpha (globulin) inhibitor, H3 polypeptide
10	ribosomal protein L5
11	ret finger protein 2
12	Homo sapiens MAIL mRNA, complete cds
13	calumenin
14	MAD (mothers against decapentaplegic, Drosophila) homolog 5
15	protein tyrosine phosphatase, receptor type, F
16	proteoglycan 1, secretory granule
17	microfibrillar-associated protein 2
18	STAT induced STAT inhibitor 3
19	Tubulin, alpha, brain-specific
20	major histocompatibility complex, class II, DQ alpha 1
21	leucine aminopeptidase
22	calponin 3, acidic
23	mannose receptor, C type 1
24	cathepsin L
25	inositol 1,3,4-triphosphate 5/6 kinase

Table B.3: Top Genes 25 Blaveri Survival

Rank	Gene Title
1	hypothetical protein PRO1847
2	enolase 2, (gamma, neuronal)
3	KIAA0672 gene product
4	transcription factor 15 (basic helix-loop-helix)
5	zinc finger protein 266
6	oxytocin receptor
7	tubby like protein 3
8	suppressor of Ty (S.cerevisiae) 4 homolog 1
9	KIAA0410 gene product
10	glutamyl-prolyl-tRNA synthetase
11	syntaxin binding protein 1
12	Homo sapiens cDNA FLJ13303 fis, clone OVARC1001372, highly similar to Homo sapiens liprin-alpha4 mRNA
13	Homo sapiens, clone IMAGE:3940519, mRNA, partial cds
14	BCL2/adenovirus E1B 19kD-interacting protein 1
15	Rag D protein
16	alanyl-tRNA synthetase
17	KIAA0027 protein

18	proteasome (prosome, macropain) subunit, beta type, 1
19	guanine nucleotide binding protein 4
20	mitochondrial ribosomal protein L12
21	chromosome 2 open reading frame 8
22	Homo sapiens cDNA FLJ10447 fis, clone NT2RP1000851
23	KIAA0981 protein
24	stage
25	grade

Table B.4: Top Genes 25 Kim Stage

Rank	Gene Title
1	dynamin 1
2	extra spindle pole bodies homolog 1 (S. cerevisiae)
3	cytoskeleton associated protein 2-like
4	defective in sister chromatid cohesion 1 homolog (S. cerevisiae)
5	chromosome 17 open reading frame 53
6	trophinin associated protein (tastin)
7	IQ motif containing GTPase activating protein 3
8	citron (rho-interacting, serine/threonine kinase 21)
9	cysteine-rich protein 1 (intestinal)
10	E2F transcription factor 1
11	CDC45 cell division cycle 45-like (S. cerevisiae)
12	centromere protein A
13	nucleolar and spindle associated protein 1
14	cyclin-dependent kinase inhibitor 3
15	monocyte to macrophage differentiation-associated
16	cell division cycle 20 homolog (S. cerevisiae)
17	aurora kinase A; aurora kinase A pseudogene 1
18	non-SMC condensin I complex, subunit G
19	cyclin B2
20	Holliday junction recognition protein
21	centrosomal protein 55kDa
22	chromosome 1 open reading frame 175
23	pyridine nucleotide-disulphide oxidoreductase domain 2
24	chromosome 8 open reading frame 16
25	interferon, epsilon

Table B.5: Top Genes 25 Blaveri Grade

Rank	Gene Title
1	GINS complex subunit 4 (Sld5 homolog)
2	septin 3
3	E2F transcription factor 1
4	dimethylarginine dimethylaminohydrolase 2
5	24-dehydrocholesterol reductase
6	extra spindle pole bodies homolog 1 (S. cerevisiae)
7	histone cluster 1, H1c
8	histone cluster 1, H2bk
9	eukaryotic translation initiation factor 4E binding protein 1
10	olfactory receptor, family 2, subfamily B, member 6
11	inositol(myo)-1(or 4)-monophosphatase 2
12	RecQ protein-like 4
13	defective in sister chromatid cohesion 1 homolog (S. cerevisiae)
14	BCL2-like 12 (proline rich)
15	cell division cycle associated 5
16	p53 and DNA-damage regulated 1
17	aurora kinase A; aurora kinase A pseudogene 1
18	similar to ferritin, light polypeptide; ferritin, light polypeptide
19	myelin protein zero-like 1
20	chromosome 9 open reading frame 140
21	solute carrier family 29 (nucleoside transporters), member 4; similar to solute carrier family 29 (nucleoside transporters), member 4
22	thymidine kinase 1, soluble
23	chromosome 1 open reading frame 112
24	chromosome 15 open reading frame 48
25	cell division cycle associated 8

Table B.6: Top Genes 25 Blaveri Survival

Rank	Gene Title
1	grade
2	stage
3	carbohydrate (N-acetylgalactosamine 4-0) sulfotransferase 8
4	adrenomedullin 2
5	ribosome binding protein 1 homolog 180kDa (dog)
6	cyclin N-terminal domain containing 2
7	lipase, endothelial
8	chromosome 5 open reading frame 46
9	espin
10	phosphodiesterase 6B, cGMP-specific, rod, beta

11	transmembrane protein 195
12	FAT tumor suppressor homolog 1 (Drosophila)
13	family with sequence similarity 13, member B
14	N-6 adenine-specific DNA methyltransferase 2 (putative)
15	plexin domain containing 2
16	chromosome 1 open reading frame 186
17	homeobox and leucine zipper encoding
18	chromosome 7 open reading frame 41
19	aspartylglucosaminidase
20	similar to programmed cell death 2
21	chloride channel 3
22	nuclear receptor subfamily 2, group C, member 1
23	N-acetylneuraminate pyruvate lyase 2 (putative)
24	arrestin domain containing 4
25	G protein-coupled receptor 98

Appendix C: Synthetic Data Set

A Synthetic Data Set was produced to test the new RBF Neural-Fuzzy Feature selection.

There are several factors to take in account to produce a synthetic Data Set:

- The output of the system is only linked to the relevant genes
- Add some noise to the data
- Produce genes that are correlated to relevant genes

The synthetic data set consists in 100 patients and 150 genes. From the Synthetic Data Set 150 genes were selected, 100 irrelevant and 50 relevant. The expectation is that the RBF Method selects only the 50 relevant.

The RBF-NF model was developed as described in section 5.2 and 5.4. The output values were 'encoded' into -1 and 1. The classification function of Accuracy is used as measure of performance. The results shown are the mean % of the 10 models for Accuracy in selecting relevant features.

The results of the new method based in the Entropy, Output Weights and Membership Functions applied to the Synthetic Data are shown below:

Table C.1: Accuracy in selecting relevant features

		Accuracy in selecting relevant features
5 rules	**(%)**	80.2
	Standard Deviation	3.8

Appendix D: Input-output mappings across different microarray technologies, showing the non-linear behaviour

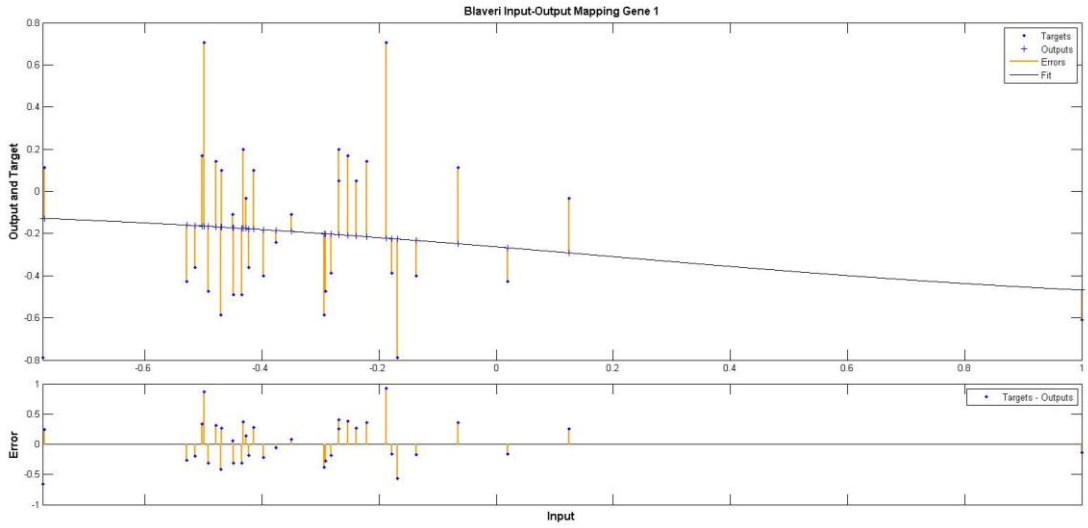

Figure D.1: Blaveri Input-Output Mapping Gene 1

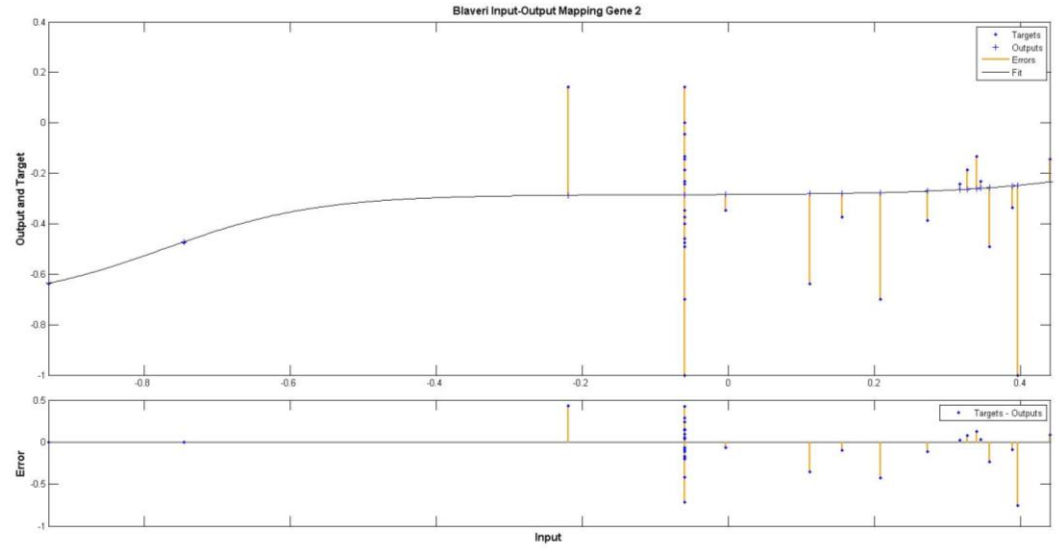

Figure D.2: Blaveri Input-Output Mapping Gene 2

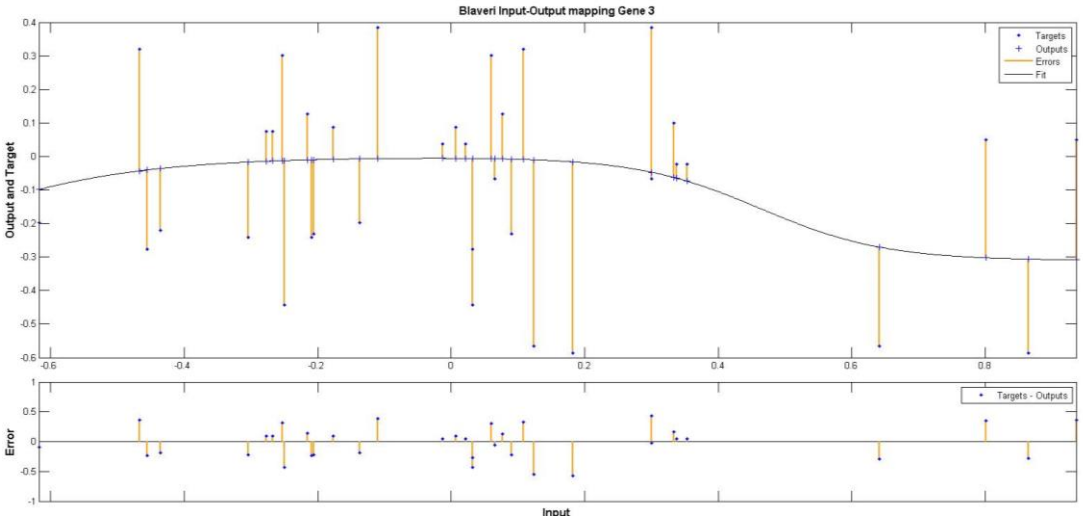

Figure D.3: Blaveri Input-Output Mapping Gene 3

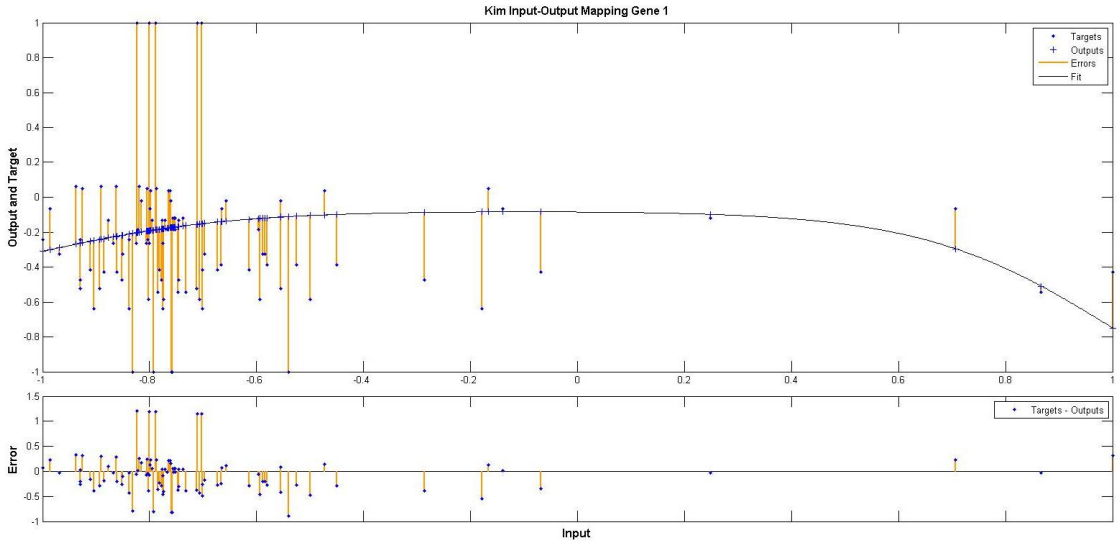

Figure D.4: Kim Input-Output Mapping Gene 1

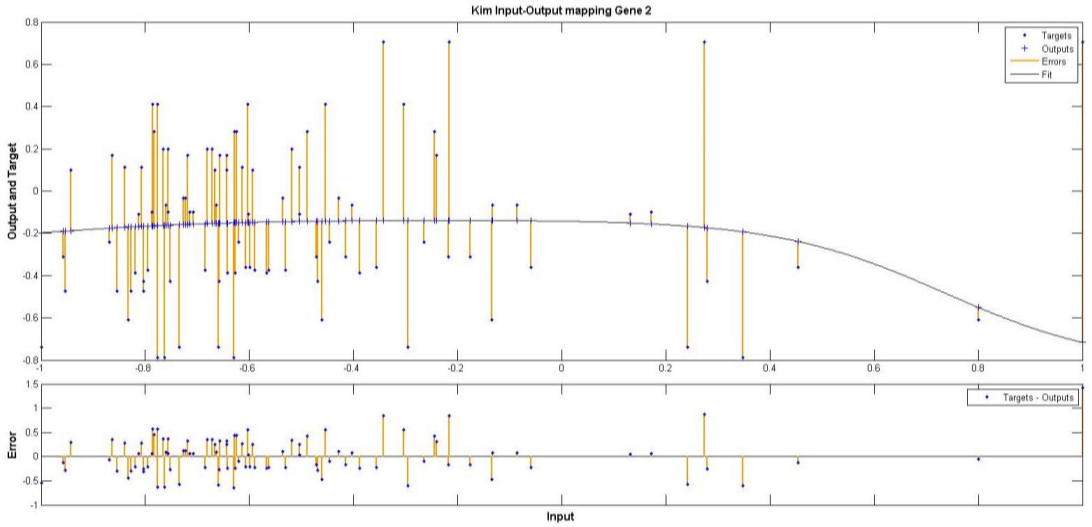

Figure D.5: Kim Input-Output Mapping Gene 2

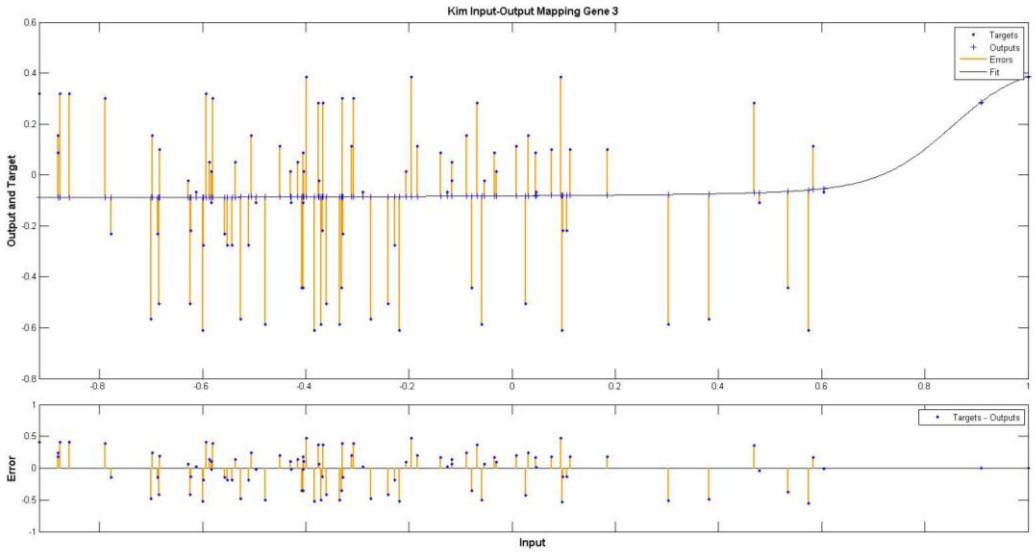

Figure D.6: Kim Input-Output Mapping Gene 3