GREY RELATIONAL ANALYSIS FEATURE SELECTION FOR CANCER CLASSIFICATION USING SUPPORT VECTOR MACHINE

SHARIFAH HAFIZAH BINTI SY AHMAD UBAIDILLAH

UNIVERSITI TEKNOLOGI MALAYSIA

GREY RELATIONAL ANALYSIS FEATURE SELECTION FOR CANCER CLASSIFICATION USING SUPPORT VECTOR MACHINE

SHARIFAH HAFIZAH BINTI SY AHMAD UBAIDILLAH

A thesis submitted in fulfillment of the
requirements for the award of the degree of
Master of Science (Computer Science)

Faculty of Computing
Universiti Teknologi Malaysia

AUGUST 2014

"This thesis is special dedicated to my adored parents, Sy Ahmad Ubaidillah and Sharifah Jamaliah, my infinitely encouraging sisters, Sharifah Nadzirah and Sharifah Hazimah and my supportive husband, Mohd Amierul Faez for their endless love and support"

ACKNOWLEDGEMENTS

All praise to Allah, the Almighty, most Gracious and most Merciful. I would like to express my sincere gratefulness to my supervisors, Assoc. Prof. Dr Roselina bt Sallehuddin for always guiding, support and inspiring me throughout this study. A million thanks for giving me endless guidance, idea and comments that improved this research until the final form.

I would like to acknowledge with much gratitude to my family for giving me morale encouragement in completing this research. My earnest appreciation also goes to Universiti Teknologi Malaysia for providing me financial support during the period of my research work. Last of all, my sincere grateful is extended to all my colleagues for giving me nonstop support and confidence during my efforts in my research as well.

ABSTRACT

Nowadays, cancer is one of the leading causes of death in the world. However, cancer can be treated if it is diagnosed earlier. Recently, machine learning classifiers are widely applied in cancer detection due to their accurate diagnosis in cancer classification problems. However, the performance of the classifiers can be affected by the selection of the required variables used in the classification process. To choose these variables, this research proposed two classification models using two different feature selection methods namely: Grey Relational Analysis (GRA) and Improved Grey Relational Analysis (IGRA). Both of these methods are combined with a Support Vector Machine (SVM) classifier and named as GRA-SVM and IGRA-SVM. The GRA and IGRA act as a feature selection method in the preprocessing phase of SVM classifier to recognize potential variables in cancer data that can be used as significant input to SVM classifier to improve SVM classification capability performance. Using performance measuring tools, the efficiency of the proposed classification models: GRA-SVM and IGRA-SVM based on the value of geometric mean, sensitivity, specificity, accuracy and area under Receiver Operating Characteristic curve were compared with standard SVM and other classification models from previous studies. The results showed that the proposed GRA-SVM and IGRA-SVM classification models have achieved better performance in classifying the cancer data with better results ranging between 2.64% to 88.9% in the selection of potential variables.

ABSTRAK

Kini, kanser adalah salah satu punca utama kematian di seluruh dunia. Namun, kanser boleh dirawat jika penyakit ini boleh didiagnosis awal. Kebelakangan ini, pengelas *Machine Learning* (ML) digunakan secara meluas dalam pengesanan kanser kerana kaedah ini dapat mendiagnosis dengan tepat dalam menyelesaikan masalah pengelasan kanser. Namun, prestasi pengelas tersebut dipengaruhi oleh pemilihan pembolehubah-pembolehubah yang diperlukan yang digunakan dalam proses pengelasan. Untuk memilih pembolehubah-pembolehubah tersebut, kajian ini mencadangkan dua model pengelasan menggunakan dua jenis teknik pemilihan ciri iaitu *Grey Relational Analysis* (GRA) dan *Improved Grey Relational Analysis* (IGRA). Kedua-dua teknik ini digabungkan dengan pengelas *Support Vector Machine* (SVM) dan dinamakan sebagai GRA-SVM dan IGRA-SVM. GRA dan IGRA bertindak sebagai teknik pemilihan ciri di dalam fasa prapemprosesan untuk mengenalpasti pembolehubah yang berpotensi yang boleh digunakan sebagai input kepada pengelas SVM bagi meningkatkan keupayaan prestasi pengelas SVM. Dengan menggunakan alat pengukuran prestasi, kecekapan dua model pengelasan yang dicadangkan iaitu GRA-SVM dan IGRA-SVM dibandingkan dengan SVM piawai dan model pengelasan yang lain daripada kajian terdahulu berdasarkan nilai min geometri, kepekaan, kekhususan, kejituan dan luas di bawah lengkungan ciri penerima operasi. Keputusan kajian menunjukkan bahawa model pengelasan GRA-SVM dan IGRA-SVM yang dicadangkan telah mencapai prestasi yang lebih baik dalam mengelaskan data kanser dengan peratusan pemilihan pembolehubah yang berpotensi di antara 2.64% ke 88.9%.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ML	Machine Learning
GRA	Grey Relational Analysis
IGRA	Improved Grey Relational; Analysis
SVM	Support Vector Machine
GRA-SVM	Grey Relational Analysis-Support Vector Machine
IGRA-SVM	Improved Grey Relational Analysis-Support Vector Machine
G-Mean	Geometric Mean
AUC	Area under Receiving Operating Characteristic Curve
ROC	Receiving Operating Characteristic
CT Scan	Computerized Axial Tomography
MRI	Magnetic Resonance Image (MRI)
ANN	Artificial Neural Network
FL	Fuzzy Logic
LDA	Linear Discriminant Analysis
DT	Decision Tree
k-NN	k-Nearest Neighbour
RF	Random Forest
NB	Naïve-Bayes
LR	Logistic Regression
CBR	Case-based Reasoning
PSO	Particle Swarm Optimization
RS	Rough Set
PCA	Principal Component Analysis

RFE	Recursive Feature Elimination
HGA	Hybrid Genetic Algorithm
LMNN	Levenberg-Marquardt Neural Network
RBFNN	Radial Basis Function Neural Network
AIRS	Artificial Immune Recognition System
SSVM	Smooth Support Vector Machine
DFA	Discrete Firefly Algorithm
CFS	Correlation based Feature Selection
GA	Genetic Algorithm
WBCD	Wisconsin Breast Cancer Dataset
RBF	Radial Basis Function
FS_SFS	Filter and Supported Sequential Forward Search
DFPA	Discriminative Function Pruning Analysis

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Cancer is one of the leading causes of death in most countries, and the number of patients survives from cancer diseases are decreasing. The survival rate is strongly influenced by the stage of the malignancy (malignant tumour) at the point of diagnosis. Therefore, a method that allows early diagnosis is desirable to increase the survival rate (Sattlecker, 2011). However, the similar appearances of some types of cancer are the main challenge for common diagnostic tools used by the medical expert such as biopsy; x-ray and MRI scan (Chen, 2011). Furthermore, the diagnosis results are subjective because they depend on the opinion of the medical experts. Thus, the results can vary even when the same sample is examined at different times either by the same medical experts or others. In addition, the total examination procedure is also time consuming (Sattlecker, 2011). Therefore, a better diagnostic approach is needed.

To improve the drawbacks of common diagnostic methods, machine learning classification method was introduced. Classification method has been widely used to solve the cancer diagnosis problem (Makinaci, 2005; Assareh, 2008; Cinar et al, 2009;

Potdukhe and Karule, 2009; Subashini et al, 2009; Chen et al, 2011; Keyvanfard et al, 2011; Ren, 2012;). In terms of cancer diagnosis, classification method is used by the researchers in order to classify tumour into two different types which are malignant tumour (cancerous) and benign tumour (non-cancerous). Classification method helps minimizing the possible errors which may occur due to inexperienced doctors. Besides, the classification method also specifies the medical data to be examined faster and more accurate (Akay, 2009). Classification methods are non-invasive, fast, low in costs, high-throughput, and only need minimal amount of training (Sattlecker, 2011).

However, for some advanced classification methods such as support vector machine, the dimension of variables vectors affects the performance of the classification and also determines the training time of the algorithm. Thus, how to extract useful variables and make a good selection of the variables is a crucial task (Osareh and Shadgar, 2009). The choice of optimal variables plays an important role in developing a classification model with high classification ability (Chen, 2011). An optimum variable set should have effective and discriminating variables, while mostly reduce the redundancy of variables pace to avoid "curse of dimensionality" problem. The "curse of dimensionality" suggests that the sampling density of the training data is too low to promise a meaningful estimation of a high dimensional classification function with the available finite number of training data (Osareh and Shadgar, 2009). With the purpose of improving the performance of the classification method, feature selection approach has been proposed. The feature selection approaches provide faster and more cost-effective classification model (Brown, 2010). Furthermore, feature selection is very useful in reducing the execution time and the dimensionality of the data to be processed by the classifier, and also improving the predictive accuracy.

1.2 Problem Background

Cancer is a type of disease that needs early treatment in order to increase the survival rate of the patient. Nowadays, there is still no single test which can accurately diagnose and determine the cancer stage. Patients who are suspected of developing a malignant tumour are examined using common diagnostic techniques such as biopsy, X-ray, MRI and CT scan (Weissleder and Pittet, 2008). However, these diagnostic techniques have some limitations such as high cost and the limited capability to identify the pattern of cancers for each patient. In order to overcome the limitations, a method that is accurate and robust, which is easier and cheaper to implement is needed. Classification method has been proposed by many researchers to determine the cancerous tumour in human body. The classification method is proven to precisely classify tumours and produce a successful diagnosis of cancer.

Classification is one of the supervised machine learning techniques: Individual item is set into classes depends on quantitative information on one or more characteristic inherent in the features and based on a set of data. The reliable performances of classification method gained a lot of attentions from the machine learning community. The results from several studies strongly suggested that the classification method performed better than the common diagnostic techniques to solve the cancer diagnosis problems. The classification method performance is supported by the study conducted by Mousa et al (2008) that used a classification method to develop a system to classify the abnormality in digital mammograms using Fuzzy and Artificial Neural Network (ANN) classifiers. Cho et al (2011) also used a classification method to analyst breast cancer mammography image by using Fuzzy Logic (FL) and Linear Discriminant Analysis (LDA) classifiers. Meanwhile, Sun et al (2013) used multiples classifier namely; Support Vector Machine (SVM), Artificial Neural Network (ANN), Decision Tree (DT), k-Nearest Neighbour (k-NN) and Random Forest (RF) to classify lung cancer

data. The results showed that classification method could produce excellent performance in diagnosis cancer.

A learning algorithm that performs classification process is known as a learner or a classifier. The classifier is split into two types; parametric classifier and non-parametric classifier (Kumar and Sahoo, 2012). The widely used parametric classifiers for cancer diagnosis are Naïve-Bayes (NB), Logistic Regression (LR) and Liner Discriminant Analysis (LDA) while the well-known non-parametric classifiers for cancer diagnosis are Artificial Neural Network (ANN), K-Nearest Neighbour (k-NN) and Support Vector Machine (SVM) (Fischer et al, 2004; Heshmati, 2011; Cınar et al, 2009; Statnikov, 2005; Kumar and Sahoo, 2012). Previous studies showed that the non-parametric classifier had outperformed the parametric classifier in terms of the percentage of correctly classified tumour. Moreover, the non-parametric classifiers can learn well and deal efficiently with high dimensional data (Enachescu, 2005; Regnier-Coudert et al, 2011).

Recently, Support Vector Machine (SVM) which is one of the commonly used non-parametric classifier had received increasing popularity in the machine learning community (Purnami et al, 2008). SVM is based on linear machine in a high dimensional feature space. It is non-linearly related to the input space which has allowed the development of fast training techniques even with a large number of input variables and big training sets (Akay, 2009). Compare to other classifier such as Artificial Neural Network (ANN) and K-Nearest Neighbour (k-NN), SVM is proven to have advantages in handling classification tasks with successful generalization performance. Other advantage of SVM is that training SVM is similar as solving a linear constrained quadratic programming problem, thus it is usually not to be trapped in the local minimum (Chen, 2011). Previous analyses have shown that SVM can give good performance in classifying various type of cancer data (Laura and Ruslan, 2008;

Akay, 2009; Cinar et al, 2009; Chen et al, 2012). Based on the advantages mentioned above, SVM is used as the classifier in this study to conduct the cancer diagnosis test.

However, for the past few years, cancer classification problems have been extensively studied. Cancer classification problems generally involve a number of variables. Not all of these varibales are equally important for a specific task. Some of them may be redundant or even irrelevant. Better performance may be achieved by discarding some variables. In other circumstances, the dimensionality of input space may be decreased to save some computation effort, although this may slightly lower classification accuracy (Lin et al, 2008). Therefore, the classification process must be fast and accurate, using the smallest number of variables. This objective can be achieved using feature selection approach. Feature selection strategies are often implied to explore the effect of irrelevant variables on the performance of classifier systems (Valentini, Muselli, & Ruffino, 2004; Zhang, Guo, Du, & Li, 2005; Acir, O¨ zdamar, & Guzelis, 2006; Akay, 2009; Chen, 2011). Furthermore, getting more information such as the most and the least significant factors that influence the cancer classifier performance is very important. However, identifying the relationship among the contributing variables is always grey particularly when the information is not clear, incomplete and uncertain. Therefore, a feature selection model that can handle the incomplete cancer data is needed.

Generally, feature selection method can be divided into two categories which are wrapper method and filter method. The wrapper methods are too expensive to be used if the number of variables is large. Filter methods are much faster and can be applied to large size datasets with many variables (Guyon et al, 2002; Huang et al, 2008; Mazlan, 2009). Grey Relational Analysis (GRA) is one of the filter feature selection methods which had been used by many researchers. It is proven to help improving the performance of classifiers (Sallehuddin, 2009; Pan and Lin, 2010; Nagpal et al, 2012;

Mat Deris et al, 2013). GRA is a multiple criteria decision support approach which develop ranking and suggest the best choice from a set of alternatives (Huang et al, 2008; Li et al, 2010). GRA has some advantages such as requiring less data, does not rely on data distribution and more applicable to a numeric data value (Zhang and Li, 2006). Besides that, GRA approach is flexible to model complex nonlinear relationships and it is proven to be an accurate and simple method for selecting factors especially for problems with unique characteristic (Sallehudin et al, 2010; Nagpal, 2012). Thus, for this study, the GRA feature selection method is used to select the potential variables and improved the performance of the SVM classifier in cancer data classification.

Despite the fact that GRA could be a good feature selection method for cancer classification, there are some limitations in the traditional GRA proposed by Deng (1989). Grey Relational Degree (GRD), used to determine the GRA ranking value, is mostly affected by the distinguishing rate between the compared sequences. Therefore, it can change substantially when the number of sequences to be compared changes. Thus, the GRD value can be fluctuating because of the slight changes in the compared sequences (Ip et al, 2009). In order to improve the performance of traditional GRA, Ip et al (2009) has proposed the improved GRA (IGRA) method. The IGRA method used the original recorded data without the need of converting them and the level of the computed variable is seen to give a better-quality grading (Ip et al, 2009). For that reason, the IGRA method approach is also used in this study as feature selection method to examine whether it can further improved the performance of the SVM classifier.

In conclusion, for the purpose of getting high diagnostic capability by using only potential variables, a study is conducted to build a new classification model that consist of a feature selection method and a classifier for cancer diagnosis. For this study, two classification models are developed using SVM as classifier with GRA and IGRA as feature selection methods. Even though there are many classification techniques that

have been introduced lately for cancer classification problems, none of them has been tested in two different types of cancer data namely standard data and gene expression data. Therefore, those techniques are not yet proven as robust techniques since none of them have been applied in both types of cancer data. Therefore, in this study, both types of cancer data are used in order to verify the robustness of the proposed classification models. Figure 1.1 summarizes the scenarios that lead to the research problem and the proposed solutions of this study.

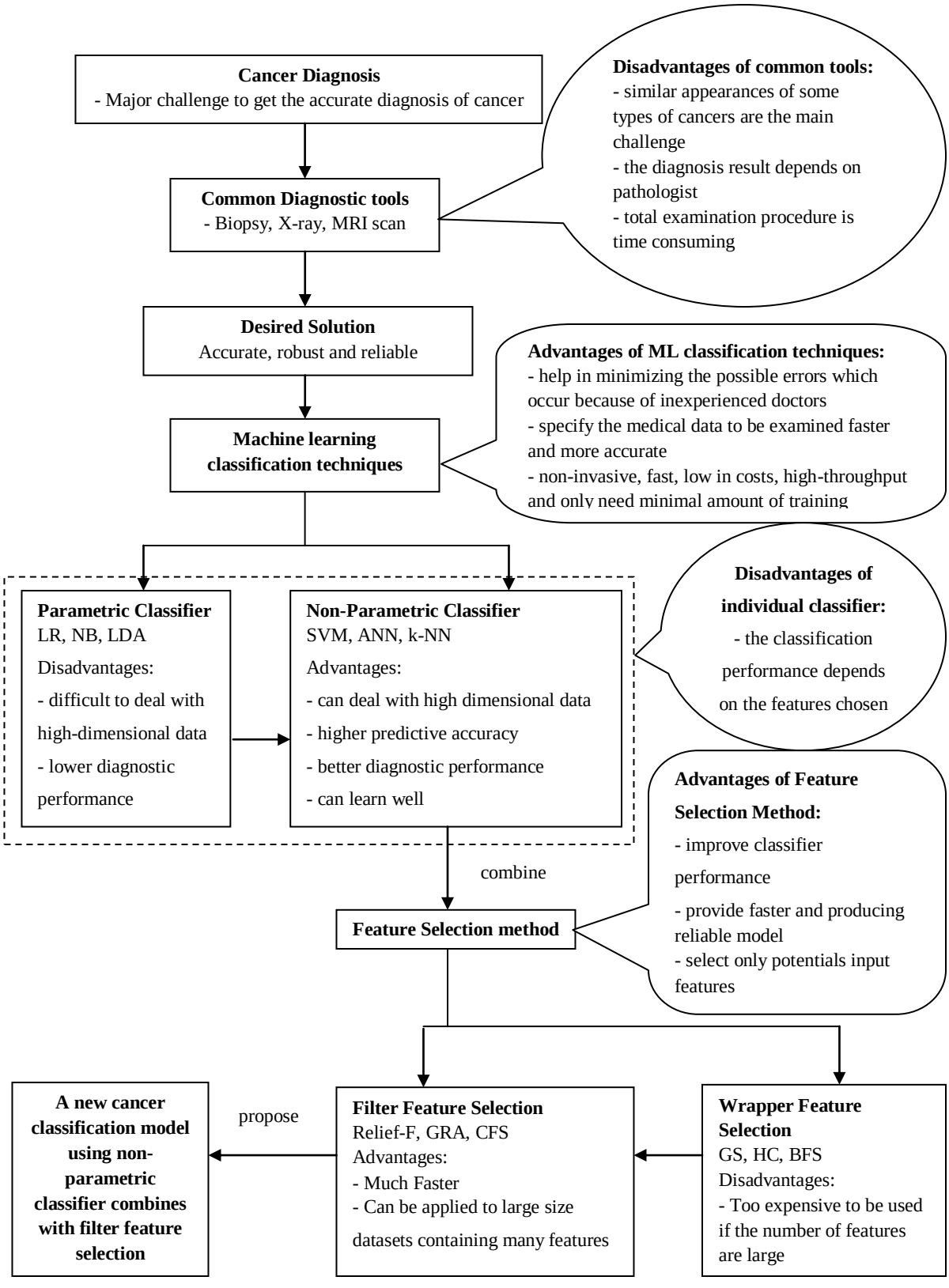

Figure 1.1 Scenario leading to the research problem and research solution

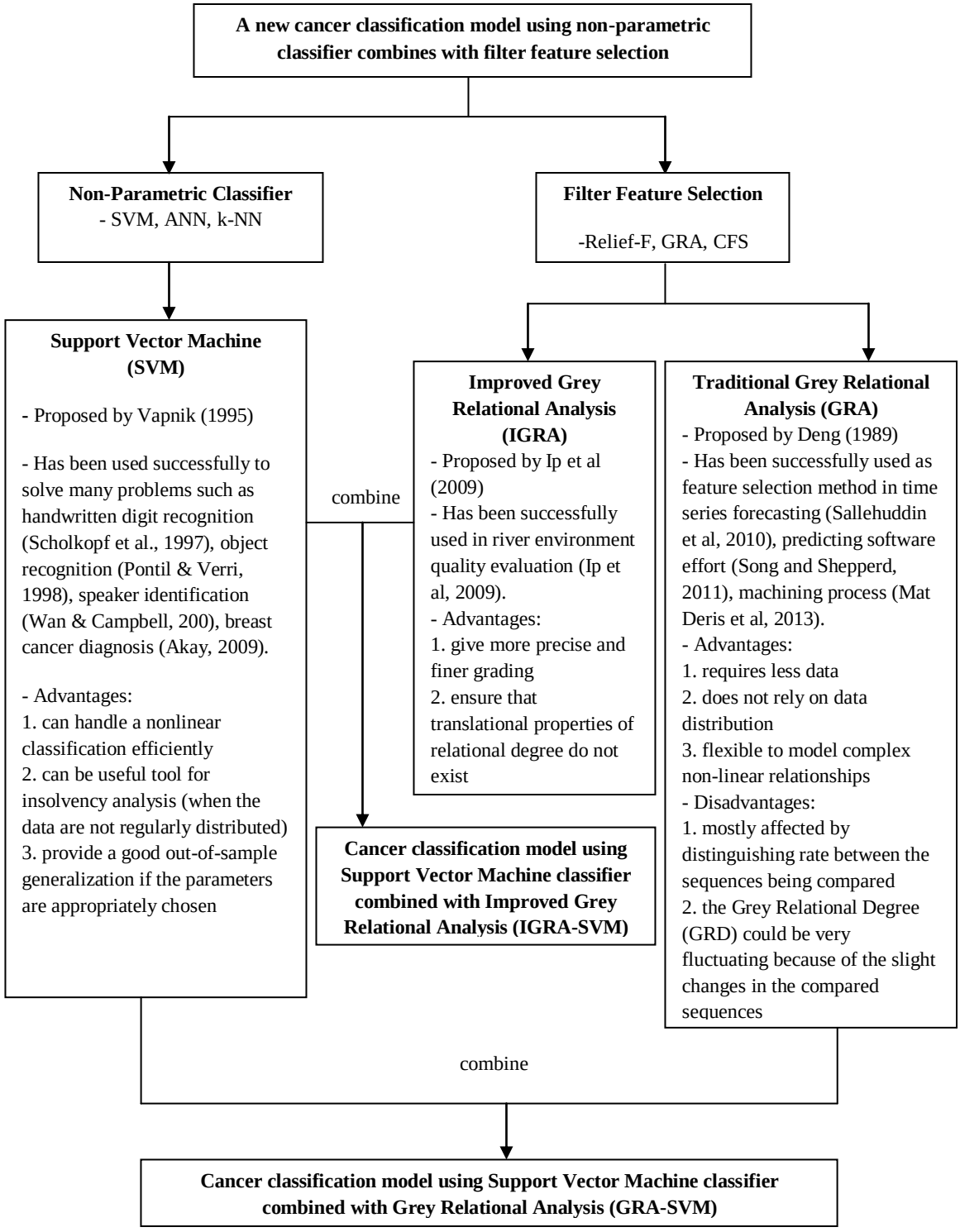

Figure 1.1 Scenarios leading to the research problem and research solution (continue)

1.3 Problem Statement

Cancer has been one of the leading causes of death in the world. However, it can be cured if it is diagnosed early. An accurate and reliable diagnosis procedure is required in early diagnosis to distinguish the benign tumour and the malignant tumour. The common diagnostic techniques that had been used by the medical expert are lacking the ability to accurately classify both tumour and are depended on the decision from the medical expert. Therefore, new diagnostic technique is needed in order to get a more accurate diagnosis on cancer.

Machine learning classification techniques are proven to provide excellent performance in cancer diagnosis. Although classification techniques are a good choice in diagnosing cancer, the performance of the classifiers can be affected if they used irrelevant and insignificant variables. For the purpose of getting an accurate diagnosis, the feature selection method is needed to remove the irrelevant and insignificant variables in order to select the only optimal variables to be processed by the classifier. However, the selection of significant variables is a big challenge in cancer classification since most of the data are incomplete

Nowadays, there are a lot of new classification model with feature selection approach has been implemented by many researchers for cancer classification. Most of them are successfully employed either in standard data or gene expression data. But none of them are applied in both types of data simultaneously. Thus, the best classification model for cancer diagnosis should be the one that is more robust, can handle incomplete data and can achieve high percentage of correctly classified tumours by using only potential variables.

Therefore the problem statement of this research is,

"A new and robust classification model with feature selection method that is able to identify the optimum variables which affect the performance of the classifier in order to get higher percentage of correctly classified tumour"

The followings are the research questions that will be addressed to answer the above problem statement:

1. How to design a new classification model?
2. How to find the least and the most important variables that affect the classifier performance?
3. How to find the optimum variables that can signify the whole pattern in cancer data in order to sustain high percentage of correctly classified tumour?
4. How to identify the efficiency of the proposed classification model?
5. Can the proposed classification model perform better than the individual models and the existing classification models with feature selection approach?

These research questions will be answered through the experimental results obtained during this study.

1.4 Research Aim

The aim of this research is to propose an accurate and robust feature selection approach for cancer classification that is able to provide better percentage of correctly classified tumour by using only potential variables which can influence the classifier performance in diagnosing cancer data.

1.5 Research Objectives

This study is conducted to achieve the above aim and the following objectives:

1. To propose Grey Relational Analysis (GRA) feature selection for cancer classification using Support Vector Machine (SVM) classifier in order to identify the potential variables to distinguish between benign tumour and malignant tumour.
2. To improve Grey Relational Analysis (GRA) feature selection by implementing Improved Grey Relational Analysis (IGRA) feature selection for cancer classification using Support Vector Machine (SVM) classifier.

1.6 Research Scope

The scope of this research is limited to the following:

1. The research only focuses on two types of cancer data which are standard cancer data (breast cancer dataset and liver cancer dataset) and gene expression cancer data (prostate cancer dataset and ovarian cancer dataset).
2. The classifier used in this research is a Support Vector Machine (SVM) classifier.
3. The approaches used for feature selection are Grey Relational Analysis (GRA) method and Improved Grey Relational Analysis (IGRA) method.

1.7 Significant of the Study

The main problem in medical diagnosis is to obtain the correct diagnosis result. Even though there are many medical experts around the world, most of them cannot give the exact and accurate diagnosis of any cancer disease. In addition, some of the commonly used diagnostic techniques are time consuming and do not have high diagnostic capability. One of the ways to solve the problems is by using machine learning classification techniques such as the proposed classification models that can help in early detection of all types of cancers and reduce the mortality rate caused by cancer. Based on the literature review, there are many classification models have been develop in order to classify cancer data. However, none of them have been tested in two different types of cancer data which are standard dataset and gene expression dataset. Thus, it can be said that the previous classification models are not robust enough because

it has no ability to classify different types of cancer data. Therefore, the proposed classification models, GRA-SVM and IGRA-SVM are developed to satisfy the needs of an accurate, robust and reliable classification model that can be applied in various types of cancer data.

1.8 Organization of the Thesis

This section describes how this thesis is organized. There are seven chapters include in this thesis. The first chapter (Introduction) described the background of the research problems and justification of the proposed feature selection approaches. The research aim, objectives, scope and significances are also expressed in this chapter. Next, Chapter 2 which is the literature review of this study presents the overview about cancer disease, the limitation of using common diagnostic tools for cancer diagnosis, the advantages of machine learning classification techniques, feature selection approach and the advantages of using feature selection approach in classification models. The selected techniques used in this study are also discussed.

The third chapter, Research Methodology, defines the operational framework of the study and describes each process or step that involve in the study. The datasets employed and the performance evaluation tools used are also discussed in this chapter. Chapter 4 gives details about the combination process of GRA-SVM classification model and the classification results obtained by the proposed GRA-SVM classification model in classifying four cancer datasets. Chapter 5 details out the explanation on development and integration of IGRA-SVM. The classification results of the proposed classification model are also presented.

Chapter 6 (Result Evaluation and Validation) gives detail on the evaluation and validation of the proposed GRA-SVM and IGRA-SVM results of four different types of cancer datasets. The evaluation process are done in two categories; the comparative performance between both proposed classification models with standard SVM classification model and the comparative performance of the proposed classification models with the previous studies. The validation process of the proposed classification models are done by using significant test. Finally, Chapter 7 (Discussion and Conclusion) discusses the findings and contributions from this study. Future directions of the research are also mentioned.

1.9 Summary

.

This chapter provides a brief introduction to the nature and contribution of the thesis. It provides the problem background, explanation of the problem domain, outlined the problem statement, the aim and objectives of this study. The scope of the study is also mentioned to set the boundary of the research. The significance of the study is also stated. Finally, thesis organisation concludes with an overview of the contents of the thesis. Next chapter presents the literature of the study, problem investigation and guideline in order to find the solution of the problem.

REFERENCES

Anthony, G., Gregg, H., & Tshilidzi, M. (2008). An SVM multi classifier approach to land cover mapping. *ASPRS Annual Conference.*

Assareh**,** A. (2008). A hybrid random subspace classifier fusion approach for protein mass spectra classification. *EvoBIO'08 Proceedings of the 6th European conference on Evolutionary computation, machine learning and data mining in bioinformatics*, pp 1-11.

Akay, M. F. (2009). Support vector machine combined with feature selection for breast cancer diagnosis. *Expert System Appl.,* 36(2), 3240-3247.

Bertsekas, D. P. (1999). Nonlinear programming.

Brown, D., N. (2010). *Breast Cancer Classification: A New Perspective from the Supervised Analysis of aCGH Data*. Master Thesis, University of Mauritius, Mauritius.

Chang, C. C., & Lin, C. J. LIBSVM: A library for support vector machine, 2001.*Software available at http://www. csie. ntu. edu. tw/~ cjlin/libsvm*.

Chang, P. C., Lin, J. J., & Liu, C. H. (2011). An attribute weight assignment and particle swarm optimization algorithm for medical database classifications. *Comput. Methods Programs Biomed*, 107(3), 382-392.

Chen, H.L., Liu, D. Y., Yang**,** B., Liu, J. & Wang, G. (2011). A new hybrid method based on local fisher discriminant analysis and support vector machines for hepatitis disease diagnosis. *Expert System Appl.,* 38(9), 11796-11803.

Chen, H. L. (2011). A support vector machine classifier with rough set-based feature selection. *Expert System Appl.,* 38(7), 9014-9022.

Chen, H. L, Yang, B, Wang, G, Wang, S. J**.**, Liu, J**.**, & Liu, D. Y. (2012). Support vector machine based diagnostic system for breast cancer using swarm intelligence. *J. Med. Syst.*, 36(4), 2505-2519.

Cho, Y. S., Chin, C. L., & Wang, K. C. (2011). Based on Fuzzy Linear Discriminant Analysis for Breast Cancer Mammography Analysis. *Conference on Technologies and Applications of Artificial Intelligence*, pp. 57 - 61.

Choi, H., Yeo, D., Kwon, S., & Kim, Y. (2011). Gene selection and prediction for cancer classification using support vector machines with a reject option. *Computational Statistics and Data Analysis*, 55, 1897–1908.

Chu, F., Xie, W., & Wang, L. (2004). Gene selection and cancer classification using a fuzzy neural network, *IEEE Annual Meeting of the Fuzzy Information*, 2, 555 – 559.

Chuang, L.Y., Wu, K. C., & Yang, C. H. (2008). Hybrid Feature Selection Method using Gene Expression Data, *Ninth IEEE International Conference on Bioinformatics and BioEngineering*, pp. 100-106. .

Cinar, M., Mehmet, E., Erkan, Z. B., & Ziya, Y. A. (2009). Early prostate cancer diagnosis by using artificial neural networks and support vector machines. *Expert Systems with Applications*, 36, 6357-6361.

Deng, J. L. (1989). Control problems of grey systems. *Systems and Control Letters*, 1(5), 288-294.

Enachescu, D., & Enachescu, C. (2005). Learning vector quantization for breast cancer prediction. In *portuguese conference on Artificial intelligence*, pp. 177-180

Evgeniou, T., Pontil, M., Papageorgiou, C., & Poggio, T. (2003). Image representation and feature selection for multimedia database search, *IEEE Trans. Knowledge Data Eng*., 15 (4), 911–920.

Fischer, E. A., Lo, J. Y., & Markey, M. K. (2004). Bayesian networks of BI-RADS™ descriptors for breast lesion classification. *26th Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, 2, pp. 3031-3034.

Guyon, I., Weston, J., & Barnhill, S. (2002). Gene Selection for Cancer Classification using Support Vector Machines. *Machine Learning*, 46, 389-422.

Hall, M. (1999). *Correlation-based feature selection for machine learning*. Doctoral Philosophy, Waikato University, New Zealand.

Hasan, H., & Tahir, N. M. (2010). Feature selection of breast cancer based on Principal Component Analysis. In *6th International Colloquium on Signal Processing and Its Applications,* pp. 1-4.

Heshmati, A., Amjadifard, R., & Shanbehzadeh, J. (2011). ReliefF-based feature selection for automatic tumor classification of mammogram images. In *Machine Vision and Image Processing (MVIP),* pp. 1-5.

Holland, J. H. (1975). *Adaptation in natural and artificial systems: An introductory analysis with applications to biology, control, and artificial intelligence*. U Michigan Press.

Hsu, H. H., Hsieh, C. W., & Lu, M. D. (2011). Hybrid feature selection by combining filters and wrappers. *Expert Systems with Applications, 38*(7), 8144-8150.

Huang, T. M., & Kecman, V. (2005). Gene extraction for cancer diagnosis by support vector machines--an improvement. *Artificial Intelligence in Medicine*, 35, 185-194.

Huang, C. L., Liao, H. C., & Chen, M. C. (2008). Prediction model building and feature selection with support vector machines in breast cancer diagnosis. *Expert Systems with Applications*, *34*(1), 578-587.

Huang, J., Cai, Y., & Xu, X. (2007). A hybrid genetic algorithm for feature selection wrapper based on mutual information. *Pattern Recognition Letters*, 28, 1825–1844.

Ip, W. C., Hu, B. Q., Wong, H., & Xia, J. (2009). Applications of grey relational method to river environment quality evaluation in China. *Journal of Hydrology*, 379(3), 284-290.

Jaganathan, P., Rajkumar, N., & Kuppuchamy, R. (2013). A Comparative Study of Improved F-Score with Support Vector Machine and RBF Network for Breast Cancer Classification. *International Journal of Machine Learning and Computing*, 2 (6), 741-745.

Jong, K., Marchiori, E., Sebag, M., & Van Der Vaart, A. (2004). Feature selection in proteomic pattern data with support vector machines. *IEEE Symposium on Computational Intelligence in Bioinformatics and Computational Biology,* pp. 41-48.

Kang, P., & Cho, S. (2006). EUS SVMs: Ensemble of Under-Sampled SVMs for Data Imbalance Problems. *Neural Information Processing*, 4232, 837-846.

Kannan, S. S., & Ramaraj, N. (2010). A novel hybrid feature selection via Symmetrical Uncertainty ranking based local memetic search algorithm. *Knowledge-Based Systems*, 23, 580-585.

Kasabov, N. (2002). *Evolving connectionist systems: Methods and applications in bioinformatics, brain study and intelligent machines*. Springer.

Keyvanfard, F., Shoorehdeli, M. A., & Teshnehlab, M. (2011). Feature selection and classification of breast cancer on dynamic magnetic resonance imaging using ANN and SVM. *American Journal of Biomedical Engineering* 1, 20-25.

Kira, K., & Rendell, L. A. (1992). A practical approach to feature selection. In *Proceedings of the ninth international workshop on Machine learning* (pp. 249-256). Morgan Kaufmann Publishers Inc..

Kononenko, I. (1994). Estimating attributes: analysis and extensions of RELIEF. In *Machine Learning: ECML-94* (pp. 171-182). Springer Berlin Heidelberg.

Kumar, Y., & Sahoo, G. (2012). Analysis of Parametric & Non Parametric Classifiers for Classification Technique using WEKA. *I.J. Information Technology and Computer Science*, 7, 43-49.

Laura A. and Rouslan A. (2008). Support Vector Machines (SVM) as a technique for solvency analysis.

Li, F., Lung, T., & Yeh, C. (2010). Comparison of filter approaches based on RVFL classifier. *Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD),* 4, pp. 1520-1524

Lisboa, P. J., & Taktak, A. F. (2006). The use of artificial neural networks in decision support in cancer: a systematic review. *Neural networks*, *19*(4), 408-415.

Liu, Y. (2009). Feature extraction and dimensionality reduction for mass spectrometry data. *Computers in Biology and Medicine*, *39*(9), 818-823.

Liu, Y., & Zheng, Y. F. (2006). FS_SFS: A novel feature selection method for support vector machines. *Pattern recognition*, *39*(7), 1333-1345.

Liu, Y., An, A., & Huang, X. (2006). Boosting prediction accuracy on imbalanced datasets with SVM ensembles. In *Advances in Knowledge Discovery and Data Mining* (pp. 107-118). Springer Berlin Heidelberg.

Luo, W., Wang, L., & Sun, J. (2009). Feature Selection for Cancer Classification Based on Support Vector Machine. *WRI Global Congress on Intelligent Systems,* 4, pp. 422-426.

Makinacı, M. (2005). Support vector machine approach for classification of cancerous prostate regions. *World Academy of Science, Engineering and Technology*, 166-169.

Mao, K. Z. (2004). Feature subset selection for support vector machines through discriminative function pruning analysis, IEEE Trans. Syst. Man, Cybern. 34(1), 60–67.

Mat Deris, A., Mohd Zain, A., & Sallehuddin, R. (2013). Hybrid GR-SVM for prediction of surface roughness in abrasive water jet machining. *Meccanica*. 1937-1945.

Mazlan, U. H., & Saad, P. (2012). Classification of breast cancer microarray data using Radial Basis Function Network. In *International Conference on Statistics in Science, Business, and Engineering (ICSSBE)*, pp. 1-4.

Meng, Y. (2006). A swarm intelligence based algorithm for proteomic pattern detection of ovarian cancer. *IEEE Symposium on Computational Intelligence and Bioinformatics and Computational Biology*, pp. 1-7.

Mousa, R., Munib, Q., & Moussa, A. (2005). Breast cancer diagnosis system based on wavelet analysis and fuzzy-neural. *Expert systems with Applications*,*28*(4), 713-723.

Nagpal, G., Uddin, M., & Kaur, A. (2012). A Hybrid Technique using Grey Relational Analysis and Regression for Software Effort Estimation using Feature Selection. *International Journal of Soft Computing and Engineering (IJSCE), ISSN*, 2231-2307.

Ngai, E. W. T., Hu, Y., Wong, Y. H., Chen, Y., & Sun, X. (2011). The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature. *Decision Support Systems*, *50*(3), 559-569.

Pan, S. M., & Lin, C. H. (2010). Fractal features classification for liver biopsy images using neural network-based classifier. In *International Symposium on Computer Communication Control and Automation (3CA)*, 2, pp. 227-230.

Polat, K., & Güneş, S. (2007). Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing*, *17*(4), 694-701.

Polat, K., & Güneş, S. (2008). Artificial immune recognition system with fuzzy resource allocation mechanism classifier, principal component analysis and FFT method based new hybrid automated identification system for

classification of EEG signals. *Expert Systems with Applications*, *34*(3), 2039-2048.

Pontil M, Verri A (1998) Support vector machines for 3-D object recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20, 637-646.

Potdukhe, M. R., & Karule, P. T. (2009). MLP NN based DSS for analysis of ultrasonic liver image and diagnosis of liver disease. *Second International Conference on Emerging Trends in Engineering and Technology*, pp. 67-72.

Purnami, S. W., Rahayu, S. P., & Embong, A. (2008). Feature selection and classification of breast cancer diagnosis based on support vector machines. In *International Symposium on Information Technology*, 1, pp. 1-6.

Regnier-Coudert, O., McCall, J., Lothian, R., Lam, T., McClinton, S., & N'Dow, J. (2012). Machine learning for improved pathological staging of prostate cancer: a performance comparison on a range of classifiers. *Artificial Intelligence in Medicine*, *55*(1), 25-35.

Ren, J. (2012). ANN vs. SVM: Which one performs better in classification of MCCs in mammogram imaging. *Knowledge-Based Systems,* 26, 144-153.

Russell, S. J., Norvig, P., Canny, J. F., Malik, J. M., & Edwards, D. D. (2003). *Artificial intelligence: a modern approach* (Vol. 74). Englewood Cliffs: Prentice hall.

Sallehudin, R. (2009). Grey relational with BP_PSO for time series foreasting. *IEEE International Conference on Systems, Man and Cybernetics* pp. 4895-4900.

Sallehuddin, R., & Shamsuddin, S. M. (2009). Hybrid grey relational artificial neural network and auto regressive integrated moving average model for forecasting time-series data. *Applied Artificial Intelligence*, *23*(5), 443-486.

Sallehuddin, R., Shamsuddin, S. M., & Hashim, S. Z. M. (2010). Forecasting small data set using hybrid cooperative feature selection. In *12th International Conference on Computer Modelling and Simulation (UKSim),* pp. 80-85.

Samui, P., & Dixon, B. (2012). Application of support vector machine and relevance vector machine to determine evaporative losses in reservoirs.*Hydrological Processes*, *26*(9), 1361-1369.

Sattlecker M., "Optimisation of Machine Learning Methods for Cancer Diagnostics using Vibrational Spectroscopy", PhD Thesis: Cranfield University, 2011.

Scholkopf B, Kah-Kay S, Burges CJ, Girosi F, Niyogi P, Poggio T (1997) Comparing support vector machines with gaussian kernels to radial basis function classifiers. IEEE Transactions on Signal Processing 45(11), 2758-2765.

Setiono, R. (2000). Generating concise and accurate classification rules for breast cancer diagnosis. *Artificial Intelligence in medicine*, *18*(3), 205-219.

Song, Q., & Shepperd, M. (2011). Predicting software project effort: A grey relational analysis based method. *Expert Systems with Applications*, *38*(6), 7302-7316.

de Souza, J. T. (2004). *Feature selection with a general hybrid algorithm*. Doctoral dissertation, University of Ottawa.

Srivastava, A., Chakrabarti, S., Das S., Ghosh, S., & Jayaraman, V. K. (2013). Hybrid Firefly Based Simultaneous Gene Selection and Cancer Classification Using Support Vector Machines and Random Forests. *Advances in Intelligent Systems and Computing*, 201, 485-494.

Statnikov, A. (2005). *Automatic cancer diagnostic decision support system for gene expression domain*. Doctoral dissertation, Vanderbilt University.

Subashini, T. S., Ramalingam, V., & Palanivel, S. (2009). Breast mass classification based on cytological patterns using RBFNN and SVM. *Expert Systems with Applications*, *36*(3), 5284-5290.

Tan, T. Z., Quek, C., Ng, G. S., & Razvi, K. (2008). Ovarian cancer diagnosis with complementary learning fuzzy neural network. *Artificial intelligence in medicine*, *43*(3), 207-222.

Vapnik, V. (1995). *The nature of statistical learning theory*. Springer.

Vu, T. N., Ohn, S. Y., & Kim, C. W. (2007). RISC: A New Filter Approach for Feature Selection from Proteomic Data. *Medical Biometrics*, 4901, 17-24.

Wahid, C. (2011). *Microarray gene selection for cancer classification using support vector machine*. Doctoral dissertation, Faculty of Arts, Business, Informatics and Education, CQUniversity Australia.

Wan V, Campbell WM (2000) Support Vector Machines for Speaker Verification and Identification. *IEEE Workshop Neural Networks for Signal Processing*, 775-784.

Weissleder, R., & Pittet, M. J. (2008). Imaging in the era of molecular oncology. *Nature*, *452*(7187), 580-589.

Wu, Y., Wang, N., Zhang, H., Qin, L., Yan, Z., & Wu, Y. (2010, August). Application of artificial neural networks in the diagnosis of lung cancer by computed tomography. In *Sixth International Conference on Natural Computation*, 1, pp. 147-153.

Xuerui, T., & Yuguang, L. (2004). Using grey relational analysis to analyze the medical data. *Kybernetes*, *33*(2), 355-362.

Yang, C. S., Chuang, L. Y., Ke, C. H. & Yang, C. H. (2008). A hybrid feature selection method for microarray classification, *IAENG International Journal of Computer Science,* 35(3).

Zhang, L. J., & Li, Z. J. (2006). Gene selection for classifying microarray data using grey relation analysis. In *Discovery Science* (pp. 378-382). Springer Berlin Heidelberg.