# BIOACTIVITY CLASSIFICATION OF ANTI AIDS COMPOUNDS USING NEURAL NETWORK AND SUPPORT VECTOR MACHINE: A COMPARISON

RAHAYU BINTI A. HAMID

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

Faculty Of Computer Science And Information System
University Of Technology Malaysia

OCTOBER 2004

*To my husband, thank you for your love and support.*
*To my son, you mean everything to me.*
*To my mother, thank you for always being there for me, supporting me and encouraging me to be the best that I can be.*

## ACKNOWLEDGEMENT

Praises to Allah for giving me the patience, strength and will to go through and complete my study. I would like to express my appreciation to my supervisor, Associate Professor Dr. Naomie bte Salim, for her support and guidance during the course of this study and the writing of the thesis. A special thanks is due to Associate Professor Dr. Siti Mariyam bte Shamsuddin for her helpful advice and insight in this study. I would also like to extend my thanks to fellow classmates who have given me the encouragement and support when I needed it. Finally, I would like to dedicate this thesis to my family. Without their love and support I would have never come this far.

# ABSTRACT

High Throughput Screening has been used in drug discovery to screen large numbers of potential compounds against a biological target by making it possible to screen tens of thousands to hundreds of thousands of compounds at the early stage of drug design. However, it is impractical to test every available compound against every biological target. Classification is an approach in classifying the compounds into active and inactive based on already known actives. In this study, Neural Network and Support Vector Machines (SVM) are used to classify AIDS data represented as 2D descriptors. Selection of compounds used is based on the most diverse compounds. The classification models will be tested using different ratios of the data set to identify whether the size of data would affect the rate of classification. Besides that, the study also analyses the effects of dimensional reduction towards the results of the two techniques. Final results indicate that SVM produces better classification results for both the original data and the reduced dimension data.

# ABSTRAK

Penggunaan High Throughput Screening untuk menyaring sejumlah besar molekul kimia yang berpotensi terhadap sasaran biologi telah memungkinkan ratusan ribuan molekul kimia dikenalpasti pada peringkat awal dalam proses penghasilan ubat. Namun, pengujian setiap molekul kimia ke atas setiap sasaran biologi adalah tidak praktikal. Kajian ini mengaplikasikan teknik Rangkaian Neural Network dan Support Vector Machines (SVM) bagi mengkelaskan data AIDS yang berbentuk 2D. Pemilihan molekul kimia yang digunakan adalah berdasarkan kepada sifat ketaksamaan yang paling tinggi. Model pengkelasan diuji menggunakan data set dengan nisbah berbeza untuk mengenalpasti kesan saiz data terhadap keputusan pengkelasan. Selain daripada itu, kajian juga menganalisa kesan pengurangan dimensi data terhadap keputusan kedua-dua teknik. Hasil keputusan kajian menunjukkan bahawa teknik SVM menghasilkan keputusan yang lebih baik bagi data asal dan juga data yang telah dikurangkan dimensinya.

# TABLE OF CONTENT


| CHAPTER | CONTENT | PAGE |
|---|---|---|
| | TITLE | i |
| | DECLARATION | ii |
| | DEDICATION | iii |
| | ACKNOWLEDGEMENT | iv |
| | ABSTRACT | v |
| | ABSTRAK | vi |
| | TABLE OF CONTENT | vii |
| | LIST OF TABLES | xii |
| | LIST OF FIGURES | xv |
| | LIST OF SYMBOLS | xviii |
| | LIST OF ABBREVIATIONS | xix |
| | LIST OF APPENDICES | xx |
| 1 | INTRODUCTION | 1 |
| | 1.0 Introduction | 1 |
| | 1.1 Problem Background | 2 |
| | 1.2 Problem Statement | 4 |
| | 1.3 Aim | 5 |
| | 1.4 Objectives | 5 |

1.5 Scope 6
1.6 Outline Of Thesis 6
1.7 Summary 7

2 LITERATURE REVIEW 8

2.0 Introduction 8
2.1 Chemoinformatics 8
2.2 The Drug Discovery And Development Process 10
2.2.1 Assay Development 10
2.2.2 Lead Identification 11
2.2.3 Lead Optimisation 11
2.2.4 Clinical Trial 11
2.2.5 Bringing The Drug Into The Market 12
2.3 Lead Identification – The Past, Present And Future 12
2.4 AIDS Data Set 15
2.5 Structure Of Chemical Molecules 16
2.6 Dimensional Reduction Techniques 18
2.6.1 Principal Components Analysis (PCA) 18
2.6.1.1 PCA Method 19
2.6.2 Factor Analysis (FA) 22
2.7 Data Mining 22
2.7.1 Data Mining Techniques 23
2.7.1.1 Predictive Modelling 23
2.7.1.2 Database Segmentation 24
2.7.1.3 Link Analysis 24
2.7.1.4 Deviation Detection 25
2.8 Classification 25
2.8.1 Linear Discriminants 26
2.8.2 *k*-Nearest Neighbour 27
2.8.3 Decision Tree 27
2.8.4 Logistic Regression 29

| | | |
|---|---|---|
| | 2.8.5 Generalize Additive Model (GAM) | 29 |
| 2.9 | Classification Methods in Chemoinformatics | 30 |
| 2.10 | Neural Network | 35 |
| | 2.10.1 Back Propagation Neural Network | 36 |
| | 2.10.2 Structuring The Network | 38 |
| | 2.10.2.1 The Input Layer | 38 |
| | 2.10.2.2 The Hidden Layer | 38 |
| | 2.10.2.3 The Number Of Node In The Hidden Layer | 39 |
| | 2.10.2.4 The Output Layer | 40 |
| | 2.10.2.5 Selecting An Activation Function | 40 |
| | 2.10.3 Advantages And Drawbacks Of Neural Network | 41 |
| 2.11 | Neural Network In Chemoinformatics | 41 |
| 2.12 | Support Vector Machine (SVM) | 43 |
| | 2.12.1 Linear SVM | 45 |
| | 2.12.2 Non Linear SVM | 48 |
| | 2.12.3 Advantages And Drawbacks Of Support Vector Machine | 50 |
| 2.13 | SVM In Chemoinformatics | 51 |
| 2.14 | Summary | 53 |
| **3** | **PROJECT METHODOLOGY** | 55 |
| 3.0 | Introduction | 55 |
| 3.1 | Problem Identification | 55 |
| 3.2 | Literature Review | 56 |
| 3.3 | Data Acquisition And Pre-Processing | 56 |
| | 3.3.1 Data Acquisition | 57 |
| | 3.3.2 Data Transformation | 58 |
| | 3.3.3 Data Labelling/Classification | 59 |

| | | | |
|---|---|---|---|
| | 3.3.4 | Dimensional Reduction | 59 |
| | 3.3.5 | Normalization Of Principal Components | 60 |
| | 3.3.6 | Data Splitting | 60 |
| 3.4 | Implementation | | 62 |
| | 3.4.1 | Back Propagation Neural Network | 62 |
| | | 3.4.1.1 Building The Network Structure | 62 |
| | | 3.4.1.2 Learning Process | 64 |
| | | 3.4.1.3 Testing Process/Output Generation | 64 |
| | 3.4.2 | Support Vector Machine (SVM) | 65 |
| 3.5 | Analysis Of Classification Performance | | 67 |
| 3.6 | Summary | | 67 |
| **4** | **RESULTS AND DISCUSSION** | | 68 |
| 4.0 | Introduction | | 68 |
| 4.1 | Classification Results On The Original Aids Data Set | | 69 |
| | 4.1.1 | Result Of Back Propagation Neural Network | 69 |
| | 4.1.2 | Result Of SVM | 75 |
| 4.2 | Classification Results On The PCA Data Set | | 87 |
| | 4.2.1 | Result Of Back Propagation Neural Network | 87 |
| | 4.2.2 | Result Of SVM | 93 |
| 4.3 | Comparison Of Classification Results Between The Original Aids Data And PCA Data | | 105 |
| 4.4 | Summary | | 107 |
| **5** | **CONCLUSION** | | 108 |
| 5.0 | Introduction | | 108 |

| | | |
|---|---|---|
| 5.1 | Findings | 108 |
| 5.2 | Advantages of Study | 109 |
| 5.3 | Contribution of Study | 110 |
| 5.4 | Conclusion | 110 |
| 5.5 | For Future Work | 111 |

**BIBLIOGRAPHY** 112

**APPENDIX** 120

# LIST OF TABLES


| TABLE NO. | TITLE | PAGE |
|---|---|---|
| 3.1 | Principal Components selection criteria | 60 |
| 3.2 | Number of sample in training and testing set | 62 |
| 3.3 | Number of node in the network layer | 63 |
| 4.1 | The best network structure for the 20:80 Aids data set | 70 |
| 4.2 | The best network structure for the 50:50 Aids data set | 72 |
| 4.3 | The best network structure for the 80:20 Aids data set | 73 |
| 4.4 | Comparison of results for all three Aids data set using Back propagation neural network | 75 |
| 4.5 | Confusion matrix for classification of 20:80 Aids data set using Linear Kernel | 77 |
| 4.6 | Confusion matrix for classification of 20:80 Aids data set using Polynomial Kernel | 78 |
| 4.7 | Confusion matrix for classification of 20:80 Aids data set using RBF Kernel | 79 |
| 4.8 | Confusion matrix for classification of 50:50 Aids data set using Linear Kernel | 80 |
| 4.9 | Confusion matrix for classification of 50:50 Aids data set using Polynomial Kernel | 81 |
| 4.10 | Confusion matrix for classification of 50:50 Aids data set using RBF Kernel | 82 |
| 4.11 | Confusion matrix for classification of 80:20 Aids data set using Linear Kernel | 83 |

| | | |
|---|---|---|
| 4.12 | Confusion matrix for classification of 80:20 Aids data set using Polynomial Kernel | 84 |
| 4.13 | Confusion matrix for classification of 80:20 Aids data set using RBF Kernel | 85 |
| 4.14 | Comparison of results for all three Aids data set using three different kernels in SVM | 86 |
| 4.15 | The best network structure for the 20:80 PCA data set | 88 |
| 4.16 | The best network structure for the 50:50 PCA data set | 90 |
| 4.17 | The best network structure for the 80:20 PCA data set | 91 |
| 4.18 | Comparison of results for all three PCA data set using Back propagation neural network | 93 |
| 4.19 | Confusion matrix for classification of 20:80 PCA data set using Linear Kernel | 94 |
| 4.20 | Confusion matrix for classification of 20:80 PCA data set using Polynomial Kernel | 95 |
| 4.21 | Confusion matrix for classification of 20:80 PCA data set using RBF Kernel | 96 |
| 4.22 | Confusion matrix for classification of 50:50 PCA data set using Linear Kernel | 97 |
| 4.23 | Confusion matrix for classification of 50:50 PCA data set using Polynomial Kernel | 98 |
| 4.24 | Confusion matrix for classification of 50:50 PCA data set using RBF Kernel | 99 |
| 4.25 | Confusion matrix for classification of 80:20 PCA data set using Linear Kernel | 101 |
| 4.26 | Confusion matrix for classification of 80:20 PCA data set using Polynomial Kernel | 102 |
| 4.27 | Confusion matrix for classification of 80:20 PCA data set using RBF Kernel | 103 |
| 4.28 | Comparison of results for all three PCA data set using three different kernels in SVM | 104 |
| 4.29 | Comparison of Classification Performance for Aids data set between Back propagation neural network and SVM | 105 |

4.30 Comparison of Classification Performance for PCA data set between Back propagation neural network and SVM 105

# LIST OF FIGURES


| FIGURE NO. | TITLE | PAGE |
|---|---|---|
| 2.1 | The Drug Discovery and Development Process | 10 |
| 2.2 | An example of structural keys | 17 |
| 2.3 | An example of a decision tree | 28 |
| 2.4 | A neural network with one hidden layer | 36 |
| 2.5 | A linear support vector machine | 44 |
| 2.6 | SVM margin | 46 |
| 2.7 | SVM input and feature space | 48 |
| 3.1 | Data Acquisition and Pre-processing phase | 57 |
| 4.1 | Comparison of Desired Target Output and Actual Output for 20:80 Aids data set | 71 |
| 4.2 | Comparison of Desired Target Output and Actual Output for 50:50 Aids data set | 72 |
| 4.3 | Comparison of Desired Target Output and Actual Output for 80:20 Aids data set | 74 |
| 4.4 | Comparison of Desired Target Output and Actual Output for 20:80 Aids data set using Linear Kernel | 77 |
| 4.5 | Comparison of Desired Target Output and Actual Output for 20:80 Aids data set using Polynomial Kernel | 78 |
| 4.6 | Comparison of Desired Target Output and Actual Output for 20:80 Aids data set using RBF kernel | 79 |
| 4.7 | Comparison of Desired Target Output and Actual Output for 50:50 Aids data set using Linear Kernel | 80 |

| | | |
|---|---|---|
| 4.8 | Comparison of Desired Target Output and Actual Output for 50:50 Aids data set using Polynomial Kernel | 81 |
| 4.9 | Comparison of Desired Target Output and Actual Output for 50:50 Aids data set using RBF kernel | 82 |
| 4.10 | Comparison of Desired Target Output and Actual Output for 80:20 Aids data set using Linear Kernel | 83 |
| 4.11 | Comparison of Desired Target Output and Actual Output for 80:20 Aids data set using Polynomial Kernel | 84 |
| 4.12 | Comparison of Desired Target Output and Actual Output for 80:20 Aids data set using RBF kernel | 85 |
| 4.13 | Comparison of Desired Target Output and Actual Output for 20:80 PCA data set | 89 |
| 4.14 | Comparison of Desired Target Output and Actual Output for 50:50 PCA data set | 90 |
| 4.15 | Comparison of Desired Target Output and Actual Output for 80:20 PCA data set | 92 |
| 4.16 | Comparison of Desired Target Output and Actual Output for 20:80 PCA data set using Linear Kernel | 94 |
| 4.17 | Comparison of Desired Target Output and Actual Output for 20:80 PCA data set using Polynomial Kernel | 95 |
| 4.18 | Comparison of Desired Target Output and Actual Output for 20:80 PCA data set using RBF kernel | 96 |
| 4.19 | Comparison of Desired Target Output and Actual Output for 50:50 PCA data set using Linear Kernel | 98 |
| 4.20 | Comparison of Desired Target Output and Actual Output for 50:50 PCA data set using Polynomial Kernel | 99 |
| 4.21 | Comparison of Desired Target Output and Actual Output for 50:50 PCA data set using RBF kernel | 100 |
| 4.22 | Comparison of Desired Target Output and Actual Output for 80:20 PCA data set using Linear Kernel | 101 |
| 4.23 | Comparison of Desired Target Output and Actual Output for 80:20 PCA data set using Polynomial Kernel | 102 |
| 4.24 | Comparison of Desired Target Output and Actual Output for | 103 |

80:20 PCA data set using RBF kernel

## LIST OF SYMBOLS

| | |
|---|---|
| N | dimension of data |
| n | dimension of input data |
| $\Re$ | feature space |
| $\mathcal{H}$ | Euclidean space |
| $y \in Y$ | output and output space |
| $x \in X$ | input and input space |
| $\langle \mathbf{x} \cdot \mathbf{z} \rangle$ | inner product between **x** and **z** |
| $K(\mathbf{x}, \mathbf{z})$ | kernel $\langle \Phi(\mathbf{x}) \cdot \Phi(\mathbf{z}) \rangle$ |
| $\mathbf{w}$ | weight vector |
| $b$ | bias |
| $\alpha$ | dual variables or lagrange multipliers |
| $L$ | primal lagrangian |
| $W$ | dual lagrangian |
| $\|\cdot\|_p$ | $p$-norm |
| ln | natural logarithm |
| $e$ | base of the natural logarithm |
| log | logarithm to the base 2 |
| $\mathbf{x}'$, $\mathbf{X}'$ | transpose of vector, matrix |
| | natural, real numbers |
| $\eta$ | learning rate |
| $\alpha$ | momentum rate |
| $\delta$ | confidence |

## LIST OF ABBREVIATIONS

| | |
|---|---|
| CA | Confirmed Active |
| CART | Classification And Regression Trees |
| CM | Confirmed Moderately Active |
| CI | Confirmed Inactive |
| ERM | Empirical Risk Minimization |
| FA | Factor Analysis |
| GAM | Generalize Additive Model |
| HTS | High Throughput Screenings |
| LRM | Logistic Regression Method |
| MARS | Multivariate Adaptive Regression Splines |
| MLP | Multi Layer Perceptrons |
| MSE | Min Squared Error |
| NCI | National Cancer Institute |
| NMR | Nuclear Magnetic Resonance |
| PCA | Principal Components Analysis |
| RBF | Radial Basis Function |
| SAR | Structure–Activity Relationship |
| SRM | Structural Risk Minimization |
| SVM | Support Vector Machine |

# LIST OF APPENDIX


| APPENDIX | TITLE | PAGE |
|---|---|---|
| A1 | Gantt Chart of Project I | 120 |
| A2 | Gantt Chart of Project II | 121 |
| B | Sample Data | 122 |
| C | Calculation in Forward Propagation | 125 |
| D | Calculation in Backward Propagation | 126 |
| E1 | Output of Stage I and II for 20:80 Aids data set using Back propagation neural network | 127 |
| E2 | Output of Stage I and II for 50:50 Aids data set using Back propagation neural network | 131 |
| E3 | Output of Stage I and II for 80:20 Aids data set using Back propagation neural network | 135 |
| F1 | Output of Stage I and II for 20:80 PCA data set using Back propagation neural network | 139 |
| F2 | Output of Stage I and II for 50:50 PCA data set using Back propagation neural network | 143 |
| F3 | Output of Stage I and II for 80:20 PCA data set using Back propagation neural network | 147 |
| G1 | Classification results of 20:80 Aids data set using SVM | 151 |
| G2 | Classification results of 50:50 Aids data set using SVM | 155 |
| G3 | Classification results of 80:20 Aids data set using SVM | 159 |
| H1 | Classification results of 20:80 PCA data set using SVM | 163 |
| H2 | Classification results of 50:50 PCA data set using SVM | 167 |

H3 Classification results of 80:20 PCA data set using SVM 171

# CHAPTER 1

# INTRODUCTION

## 1.0 Introduction

The new millennium has ushered in an era of science that will revolutionize a great majority of our daily activities. Advances in Artificial Intelligence (AI) and its applications has made problem solving a much easier task. It plays a big role in the evolution of data mining by offering sophisticated techniques such as expert systems, heuristics, neural networks and support vector machines. Data mining seeks to discover hidden facts or information contained within raw data that the user could act upon, like making a prediction. A classification problem aims to identify the characteristics that indicate the group to which each case belongs. This pattern can then be used both to understand the existing data and to predict how new instances will behave.

The task of classification occurs in a wide range of fields and applications. Among of the applications are rainfall prediction (Chen *et al.*, 1993), bankruptcy prediction (Lacher *et al.*, 1995), handwriting recognition (Guyon, 1991) and medical diagnosis (Liu *et al.*, 2003; Burke, 1994). In this study, the field of interest is chemoinformatics, particularly in drug discovery.

## 1.1 Background of Problem

Over the past twenty years, the philosophy behind drug discovery has change radically. The traditional process of drug discovery involves an iterative process of finding compounds that are active against a protein target. Each iteration involves selecting compounds to react with that protein target in the desired manner. Better understanding of the reasons for activity is achieved by analyzing the resulting data in each iteration. This in turn will lead to a better design or compound selection in the next iteration.

Thousands of compounds are tested against the target each day to find out which compound are active i.e. binds to the target. The iteration is repeated until the best active compounds are found. The compounds may come from a variety of sources such as combinatorial chemistry, vendor catalogues or corporate collection. However, it is impractical to test every available compound against every biological target. Therefore, there is a great need to optimize this high throughput screening by developing methods that can identify promising compounds from a large chemical inventory on the basis of a relatively smaller set of tested compounds. One approach is to use the data from tested compounds to relate biological activity to molecular descriptors of chemical structures. A major challenge is although the data set may contain a large number of tested compounds, active compounds are often rare (An and Wang, 2001).

At any stage of the process, three types of compounds can be distinguished: (a) a very small fraction of compounds that have already been identified as active, (b) a much larger fraction of compounds that already have been identified as inactive, and (c) by far the largest fraction of compounds that have not yet been tested (the unlabeled compounds) (Warmuth *et al.*, 2003). Therefore, among the three types, only the active compounds will be considered for a clinical trial to produce a potential drug.

The need for a more refined search than simply producing and testing every single molecular combination possible has meant that statistical methods and, more recently, intelligent computation have become an integral part of the drug production process. Artificial intelligence techniques have been applied to narrow down the search and lessen the time needed in finding an active compound since the late 1980s, mainly in response to increased accuracy demands. The techniques used range from straightforward statistical classification methods, such as nearest neighbour and linear discriminant classifiers to more sophisticated methods, such as decision trees and neural networks. Unsupervised learning techniques, such as clustering and Kohonen networks are also used for data visualization and compound selection (Trotter *et al.*, 2001).

Hence, this study tries to apply neural network using back propagation algorithm and Support Vector Machine (SVM) in finding out quickly which compounds are active, i.e., binding to a particular target. Both techniques have been used in drug discovery before, but not with bit string data. Neural network is applied as it has been widely accepted to produce accurate results. Meanwhile, SVM is applied because they have a simple geometric motivation and also yield very good results. However, more investigations are required for applying SVM in cheminformatics.

Mathematically, a library with **n** compounds and represented by **m** (**m** >3) descriptors is an **n** × **m** dimensional matrix. There is no way to graph the matrix, although one would like to review the diversity graphically. In order to solve this problem, dimensionality needs to be reduced to two or three. Therefore, Principal Components Analysis (PCA) is applied to reduce data dimension. Principal component analysis (PCA) are usually used to filter out redundant descriptors and, eliminate descriptors having minor information contribution. PCA transforms a number of potentially correlated variables (descriptors) into a number of relatively independent variables that then can be ranked based upon their contributions for explaining the whole data set. The transformed variables that can explain most of the information in the data are called principal components. The components having

minor contribution to the data set may be discarded without losing too much information.

Nonetheless, its effectiveness in chemical classification is yet to proven. Hence, this study is conducted to identify which technique has the ability to produce the best results based on the type of molecular structure used, i.e. bit string. Also to investigate whether by reducing the dimensionality of data can generate better output as compared to its original data.

## 1.2 Problem Statement

The need to produce the latest effective drugs has led to the use of information technology in its development process. For every protein or virus, there are certain chemical compounds that would react to it, therefore considered against that target. Classification of compounds into active and inactive based on already known actives can eliminate compounds that have low possibilities of being active from being tested.

The current situation is, there exist a large number of compounds need to be mined for finding out quickly which compounds are active, i.e., binding to a particular target. The compounds may come from different sources such as vendor catalogs, corporate collections, or combinatorial chemistry. In fact, the compounds need to exist only virtually, being defined in terms of their descriptor vectors. However this is difficult to achieve as the number of active compounds are much smaller compared to inactive compounds and sometimes the chosen compounds do not result into a drug.

# BIBLIOGRAPHY

Alaimo, P. J.; Shogren-Knaak, M. A.; Shokat K. M. (2001). "Chemical genetic approaches for the elucidation of signaling pathways." *Journal of Current Opinion in Chemical Biology*. **5**. 360-367.

An, A. and Cercone, N. (1998). "ELEM2: A Learning System for More Accurate Classifications." *Proceeding of the 12th Canadian Conference on Artificial Intelligence*, Vancouver, Canada.

An, A. and Wang, Y. (2001). "Comparisons of Classification Methods for Screening Potential Compounds." *IEEE*. 11-18.

Back, A. D. (1999). "Classification Using Support Vector Machines," RIKEN Brain Institute, Wako-shi, Saitama, Japan.

Bailey, D. and Thompson, D. (1990). "How to Develop Neural Network Applications." AI Expert.

Blackwell, H. E.; Perez, L.; Stavenger, R. A.; Tallarico, J. A.; Cope Eatough, E. *et al.* (2001). "A one-bead, one-stock solution approach to chemical genetics: part 1." *Journal of Chemical Biolology*. **8**. 1167-1182.

Blake, C.L., Merz, C.J., 1998. UCI Repository of Machine Learning Databases, http://www.ics.uci.edu/_mlearn/MLRepository.html.

Brown, R. D. and C. Martin, Y. C. (1996). "Use of Structure-Activity Data To Compare Structure-Based Clustering Methods and Descriptors for Use in Compound Selection." *Journal of Chemical Information and Computer Science*. **36**. 572-584.

Brown, C. S. (1997). "Classification Using Neural Networks." http://www.cs.brown.edu/couses/cs141/outlines/lect22.html

Brown, F. K. (1998). "Chemoinformatics, what it is and how does it impact drug discovery." Annual Report of Medical Chemistry. **33**, 372-384.

Burbidge, R., Trotter, M.W.B., Buxton, B.F. and Holden, S.B. (2001). "Drug design by machine learning: support vector machines for pharmaceutical data analysis." *Journal of Computers and Chemistry*. **26**. 5-14.

Burges, C.J.C. (1998). "A Tutorial on Support Vector Machines for Pattern Recognition." *Data Mining and Knowledge Discovery 2*. **2**. 1-47.

Burke, H. B. (1994). "Artificial neural networks for cancer research: Outcome prediction." *Sem. Surg. Oncol.* **10**. 73–79.

Cabena, P., Hadjinian, P., Stadler, R., Verhees, J. and Zanasi, A. (1998). "Discovering Data Mining, From Concept To Implementation." Englewood Cliffs: Prentice Hall.

Can, Y.; Jackle, L. D.; Botton, L. et al. (1995). "Learning Algorithms for Classification: A Comparison on Handwritten Digit Recognition", "Neural Networks: The Statistical Mechanics Perspective" Oh, J. H., Kwon, C., Cho, S., Eds. World Scientific. 261-276.

Chang, C.-C. and C.-J. Lin (2001). "LIBSVM: a library for support vector machines". Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

Chen, Tao and Takagi, Mikio (1993). "Rainfall Prediction of Geostationary Meteorological Satellite Images Using Artificial Neural Network." Japan: University of Tokyo.

Chen, X. and Reynolds, C. H. (2002). "Performance of Similarity Measures in 2D Fragment-Based Similarity Searching: Comparison of Structural Descriptors and Similarity Coefficients." *Journal of Chemical Information and Computer Science*. **42**. 1407-1414

Chen, Y. Z.and Zhi, D. G. (2001). "Ligand-protein inverse docking and its potential use in the computer search of protein targets of a small molecule." *Proteins*. 43. 217-226.

Clark, L. A. and Pregibon, D. (1991). "Tree Based Models" in *Statistical Models in S*, edited by J. M. Chambers and T. J. Hastie. Wadsworth & Brooks/cole Advanced Books & Software. Pacific Grove, California.

Clemons, P. A.; Koehler, A. N.; Wagner, B. K.; Sprigings, T. G.; Spring, D. R. *et al.* (2001). "A one-bead, one-stock solution approach to chemical genetics: part 2." *Journal of Chemical Biology*. **8**. 1183-1195.

Darus, M., Ahmad, R. and Shamsuddin, S. M. Hj, (2000). "Activation Function of Sigmoid and Hyperbolic Tangent in Training Backpropagation Model - (A Comparison)", *Chiang Mai J. Sci*. **27**(1). 57-64.

Developmental Therapeutics Program NCI/NIH, AIDS Antiviral Screen Available Public Data at http://dtp.nci.nih.gov/docs/aids/aids_data.html

Diercks, T.; Coles, M.; Kessler, H. (2001). "Applications of NMR in drug discovery." *Current Opinion in Chemical Biology*. **5**. 285-291.

Duda, R.O., Hart, P.E. and Stork, D.G. (2000). "Pattern Classification." 2nd. ed., New York: John Wiley & Sons.

Gillet, V. (2004). "Introduction to Computational Tools for Combinatorial Chemistry." UK: University of Sheffield. Unpublished.

Guyon, I. (1991). "Applications of neural networks to character recognition," *International Journal of Pattern Recognition and Artificial Intelligence*. **5**. 353–382.

Hengl, T. (2001). "What is Data Mining? Finding the Pattern." PC AI Magazine Issue 5. **15**. 18-23,

Hicks, R. P. (2001). "Recent advances in NMR: Expanding its role in rational drug design." *Journal of Current Medical Chemistry*. **8**. 627-650.

Hsu C.-W., Chang, C.-C. and Lin, C.-J. (2003). "A Practical Guide to Support Vector Classification." National Taiwan University.

Illingworth, H.T, (1989)."Beginners Guide To Neural Networks", *IEEE* AES Magazine.

James, C. A., Weininger, D., Delany, J. (2004). "Daylight Theory Manual". Daylight Chemical Information Systems, Inc. CA.

Jastini Mohd Jamil (2003). "Pengkelasan Terhadap Data Pra-Pendiskretan Dan Pasca-Pendiskretan Menggunakan Set Kasar Dan Rambatan Balik: Satu Perbandingan". Universiti Teknologi Malaysia: Tesis Sarjana.

Joachims, T. (1997). "Text Categorization with Support Vector Machines." LS_ Techical Report, 23. University of Dortmund. ftp://ftpal.informatik.unidortmunddc /pub/report23.ps.Z.

Johnson, D. E. (1998). "Applied Multivariate Methods for Data Analysis". USA: Duxbury Press.

Kamruzzaman, J. and Aziz, S.M, (2002). "A Note on Activation Function in Multilayer Feedforward Learning", *IEEE*.

Kiang, M.Y. (2003). "A comparative assessment of classification methods," *Decision Support Systems*. **35**. 441-454.

Kurogi, Y.; Guner, O. F. (2001). "Pharmacophore modeling and three-dimensional database searching for drug design using catalyst." *Journal of Current Medical Chemistry*. **8**. 1035-1055.

Lacher, R. C., Coats, P. K., Sharma, S. C. and Fant, L. F. (1995). "A neural network for classifying the financial health of a firm," *Eur. J. Oper. Res.*, **85**. 53–65.

Lipmann, R. P. (1987). "An Introduction to Computing with Neural Nets." IEEE ASSP Magazines. 4-32.

Liu, H. X., Zhang, R. S., Luan, F., Yao, X. J., Liu, M. C., Hu, Z. D., and Fan, B. T. (2003). "Diagnosing Breast Cancer Based on Support Vector Machines." *Journal of Chemical Information and Computer Science*. **43**. 900-907.

Manly, C. J.; Louise-May, S.; Hammer, J. D. (2001), "The impact of informatics and computational chemistry on synthesis and screening." *Drug Discovery Today*. **6**. 1101-1110.

Masters, T. (1994). "Practical Neural Network Recipes In C++." New York: Academic Press.

Md Sah Hj. Salam, Dzulkifli Mohamad dan Sheikh Husain Sheikh Salleh (2000). "Pengecaman Sebutan Digit Menggunakan Rangkaian Neural: Satu Kajian Terhadap Carian Bilangan Nod Tersembunyi dan Parameter Pembelajaran." Jurnal Teknologi Maklumat.

Meyer, E. F.; Swanson, S. M.; Williams, J. A. (2000). "Molecular modelling and drug design." *Pharmacol. Ther.*, **85**. 113-121.

Michie, D., Spiegelhalter, D.J., Taylor, C.C.(1994). "Machine Learning, Neural and Statistical Classification". Englewood Cliffs, N.J.: Prentice Hall.

Neamati, N. and Barchi, J. J. Jr. (2002). "New Paradigms in Drug Design and Discovery." *Curr. Top Med. Chem.*, **2**(3). 211-239.

Neya, M. (2002). "Lead identification in drug discovery research", *The Journal of Structural Biology Sakabe Project*. Exploratory Research Laboratories, Fujisawa Pharmaceutical Co. Ltd. Japan.

Nolan, J.R. (2002). "Computer systems that learn: an empirical study of the effect of noise on the performance of three classification methods." *Expert Systems with Applications*. **23**. 39-47.

Oja, E. (1989). "Neural Networks, Principal Components and Subspaces." *International Journal of Neural System*. **1**(1). 61-68.

Osuna, E.; Freund, R.; Girosi, F. (1997). "Training support vector machines: An application to face detection." In *Proceedings of Computer Vision and Pattern Recognition*.130-136.

Quinlan, J. R. (1983). " Learning efficient classification procedures and their application to chess end games." In R. S. Michalski, J. G. Carbonell and T. M. Mitchell, "Machine Learning: An Artificial Intelligence Approach". Volume **1**. Palo Alto, CA: Tioga.

Roliana Ibrahim (2001). "Carian Corak Kelas Data Indeks Komposit BSKL Dalam Perlombongan Data Menggunakan Model Rambatan Balik." Universiti Teknologi Malaysia: Tesis Sarjana.

Roselina Salleh @ Sallehuddin (1999). "Penggunaan Model Rangkaian neural Dalam Peramalan Siri Masa Bermusim." Universiti Teknologi Malaysia: Tesis Sarjana.

Rumelhart, D., Mc Clelland, J. and PDP Research Group (1986). Parallel Distributed Processing: Explorations in the Microstructure of Cognition, Vol. 1, Chap. 8, Cambridge, Massachusetts: MIT Press.

Sarle, W. S. (1997). Neural Network FAQ. Periodic posting to the Usenet newsgroup comp.ai.neural-nets.

Schreiber, S. L. (1998). "Chemical genetics resulting from a passion for synthetic organic chemistry." *Bioorg. Med. Chem.*, **6**. 1127-1152.

Smith L. (2002). "A Tutorial on Principal Components Analysis". Students Tutorials on Computer Vision, University of Otago.

Stockwell, B. R.(a) (2000). " Chemical genetics: ligand-based discovery of gene function." *Nat. Rev. Genet.*, **1**. 116-125.

Stockwell, B. R. (b) (2000). "Frontiers in chemical genetics." *Trends Biotechnology* , **18**. 449-455.

Trotter, M.W.B., Buxton, B.F. and Holden, S.B. (2001). "Support Vector Machines in Combinatorial Chemistry." Department of Computer Science, University College London.

Two Crows Corporation (1999) "Introduction to Data Mining and Knowledge Discovery". Third Edition, Potomac, MD. Available at http://www.twocrows.com/intrdm.pdf.

Warmuth, M.K., Liao, J., Ratsch, G., Mathieson, M., Putta, S., and Lemmen, C. (2003). "Active Learning with Support Vector Machines in the Drug

Discovery Process." *Journal of Chemical Information and Computer Science*. **43**. 667-673.

Weiss, E. L.; Bishop, A. C.; Shokat, K. M.; Drubin, D. G. (2000). "Chemical genetic analysis of the budding-yeast p21-activated kinase Cla4p." *Natural Cell Biology*, **2**. 677-685.

Wild, D. J. (2004). "A Brief Introduction to Chemoinformatics". University of Michigan. Unpublished.

Wong, M. A. and Lane, T. (1983). "A kth nearest neighbor clustering procedure." *Journal of the Royal Statistical Society*, Series B 45, 362– 368.

Xu, J. and Hagler, A., (2002). "Chemoinformatics and Drug Discovery." *Molecules*. **7**. 566-600.

http://www2.chemie.uni-erlangen.de/presentation/symposium/willett/start.html