



**UNIVERSITY OF
CANBERRA**

AUSTRALIA'S CAPITAL UNIVERSITY

**Mining Health Data for Breast Cancer
Diagnosis Using Machine Learning**

Mohammad Ashraf Bani Ahmad

A thesis submitted for the requirements of the
Degree of Doctor of Philosophy
Faculty of Education, Science, Technology & Mathematics

December 2013



In the name of Allah, the Most Merciful, the Most Compassionate. Recite and your Lord is the most Generous - Who taught by the pen - Taught man that which he knew not. **Surat Al-`Alaq (The Clot), the Holy Quran.**

Abstract

The recent advancements in computer technologies and storage capabilities have produced an incredible amount of data and information from many sources such as social networks, online databases, and health information systems. Nowadays, many countries around the world are changing the way of implementing health care to the patients and the people by utilising the benefits of advancements in computer technologies and communications through electronic health.

Electronic health (eHealth) is the process of using emerging information and communication technologies in health care for the benefit of humans. eHealth includes a range of components such as electronic health records, electronic prescriptions, electronic and mobile treatments for patients. In Australia, the majority of medical and health coverage is provided by the government and due to shortage of medical personnel and appropriate supportive technologies, many people have to suffer long waiting times and limited medical resources. Therefore, the Australian government, territory, and state governments raised the inclusion of eHealth technologies in the health care system, to cope with the increased demand on health services and help solve some problems that face the traditional health systems.

This initiative produced the National eHealth Transition Authority Limited (known as NEHTA). The main purpose of NEHTA is to develop better ways of electronically collecting and securely exchanging health information across Australia. Since July 2012, anyone seeking healthcare in Australia can register for a personally controlled electronic health record. This can lead to a huge repository about Australian health care records.

This huge amount of data can be tuned into knowledge and more useful form of data by using computing and machine learning tools. It is believed that engineering this amount of data can aid in developing expert systems for decision support that can assist physicians in diagnosing and predicting some debilitating life threatening diseases such as breast cancer. Expert systems for decision support can reduce the cost, the waiting time, and free human experts (physicians) for more research, as well as reduce the errors and mistakes that can be made by humans due to fatigue and tiredness. However, the process of utilising health data effectively, involves many challenges such as the problem of missing features values, the curse of dimensionality due to a large number of features (attributes), and the course of actions to determine the features, that can lead to more accurate results (more accurate diagnosis). Effective machine learning tools can assist in early detection of diseases such as breast cancer, and the current work in this thesis focuses on investigating novel approaches to diagnose breast cancer based on machine learning tools, and involves development of new techniques to construct and process missing features values, investigate different feature selection methods, and how to employ them into diagnosis process.

It is believed that the adoption of electronic health systems into the health care system requires comprehensives design and development, which may need several stages to make it more useful for humans and governments. For example, storing health records and electronic exchange of health records across the country are not the only aims of eHealth. Treating health records as an important information resource and probing the data to extract useful diagnostic and disease related intelligence, by using automated approaches, including most significant features, for example, may lead to new tools/approaches to examine new cases (patients)

based on previous and similar cases, using machine learning and computer intelligence. It is the process of mapping the existing data into new unseen scenarios and settings, that can lead to increase in understanding the disease related information, such as early onset of disease, and better monitoring of different stages of disease, leading to value addition of health care technologies, for enhanced quality of service to patients, providing better assistance to doctors (bring an electronic consultant for doctors for example), and easy to cross validate standard disease diagnostic procedures. The thesis proposes several approaches to make this vision a reality. The main findings of this research can be categorised as follows:

- The thesis proposed a new approach for diagnosing breast cancer by reducing the number of features to the optimal number using the information gain algorithm, and then applies the new reduced features dataset to the Adaptive Neuro Fuzzy Inference system (ANFIS). It is found that the accuracy for the proposed approach is 98.24%, significantly better. The promising results may lead to further attempts to utilise and exploit information technology for diagnosing patients, and provide decision support to physicians.
- The thesis proposed a new approach for constructing missing features values based on iterative k nearest neighbours and the distance functions. The approach is an iterative approach until finding the most suitable features values that satisfy classification accuracy. The proposed approach showed improvement of 0.005 of classification accuracy on the constructed dataset than the original dataset on both Euclidean and Minkowski distance functions. The study found that Manhattan, Chebychev, and Canberra distance metrics produced lower classification accuracy on the new dataset than the original dataset. The study also noticed that classification accuracy

depends greatly on the number of neighbours (k). The experimental evaluation showed that less neighbours may lead to more accuracy. The reason for that, in my opinion, is the amount of noise produced from conflict neighbours. Finally, the maximum classification accuracy was on $k=1$ which was 0.9698.

- Different sets of experiments were performed to evaluate benchmark attributes selections methods on well-known publicly available dataset from UCI machine learning repository, Wisconsin Breast Cancer dataset (WBC). Naïve Bayes has performed the supreme in regard to classification accuracy. k -NN and Decision Tree have performed just better on dataset after applying features selection methods. In general, features selection methods can improve the performance of learning algorithms. However, no single feature selection method that best satisfy all datasets and learning algorithms.
- In regards to Classification Fusion on three well-known machine learning classifiers on breast cancer dataset. the study confirms the argument that the best combination of a set of classifiers depends on the application and on the classifier characteristics. In addition, there is no best combination of classifiers that suites all datasets. However in the current experiments, Naïve Bayes and k -NN produced better results when they combined as one classifier with maximum classification accuracy obtained on WBC dataset (0.9642).

Acknowledgements

I would like to thank everyone who has helped me to complete this thesis. Special, deep, and honest thanks to the supervision panel, chiefly, Dr Girija Chetty for the guidance, smile, and her advice through this research. Big thanks to all faculty staff and employees, I'm pleased for being a small part of such great place for more than three years, principally Professor Dharmendra Sharma, A/Prof. Dat Tran for his support provided through my journey, and Professor George Cho for his comments and suggestions.

My distinctive thanks to my wife; you were a complete package of family and love that gave me strength, support, and love during tough times. My son Hashim and daughter Yarra, you are my world, my words, my strength, and the reason of my life. Thank you.

Words can't express my thanks to my family, the source of power and strength; my parents for raising me up, especially my Mum, thanks aren't enough and wouldn't be enough, and sorry for being away from your warm and kind bosom but I'm back, hopefully soon. My Dad, thank you for everything you done to me, for the advice, encouragement, and support through my life. My brothers and sisters, thank you.

This thesis is dedicated to Mum, Dad, wife Fayha, son Hashim, and daughter Yarra.

Table of Contents

Abstract	iii
Acknowledgements	ix
Table of Contents	xi
List of Figures	xiv
List of Tables	xvi
Acronyms	xvii
Chapter One: Introduction	1
1.1 Overview.....	1
1.2 Research Motivation	3
1.3 Research Objectives.....	7
1.4 Research Contribution.....	11
1.5 Research Methodology	14
1.6 Thesis Road Map	15
1.7 Chapter Summary	17
Chapter Two: Background Study and Literature Review	18
2.1 Overview.....	18
2.2 Background Study.....	18
2.3 Classification.....	19
2.3.1 <i>k</i> -Nearest Neighbors algorithm	21
2.3.2 Artificial Neural Network	27
2.3.3 Decision Tree	31
2.3.4 Naïve Bayes Classifier	34
2.4 Data Mining in Healthcare	36
2.4.1 Treatment Effectiveness.....	36
2.4.2 Healthcare Management	37
2.4.3 Customer Relationship Management	37
2.4.4 Fraud and Abuse	37
2.4.5 Computer Aided Diagnosis.....	38
2.4.6 Ethical, Legal, and Social Issues.....	40
2.4.7 Challenges of Data Mining in Healthcare.....	43
2.4.8 Electronic Health Record.....	45
2.5 Related Work on Breast Cancer Diagnosis.....	46
2.6 Feature Selection Techniques	47

2.6.1	Wrapper Feature Selection Technique	49
2.6.2	Filters Feature Selection Techniques	51
2.6.3	Embedded Feature Selection Techniques	52
2.6.4	Feature Selection Techniques Used in Current Work.....	53
2.6.5	Related Work on Feature Selection Techniques	56
2.7	Missing Features Values	58
2.7.1	Types of Missing Values.....	59
2.7.2	Handling missing data.....	60
2.8	Chapter Summary	65
Chapter Three: Research Methodology.....		66
3.1	Introduction.....	66
3.2	Data Mining Methodology	68
3.2.1	Data Collection	70
3.2.2	Data Selection	72
3.2.3	Data Pre-Processing	72
3.2.4	Applying Data Mining Methods	73
3.2.5	Evaluation	75
3.2.6	Machine Learning Software Development Tools	76
3.2.7	Results Visualization.....	76
3.3	Chapter Summary	77
Chapter Four: Breast Cancer Diagnosis Based on Information Gain and Adaptive Neuro Fuzzy Inference System.....		78
4.1	Overview.....	78
4.2	Adaptive Neural Fuzzy Inference System (ANFIS)	78
4.2.1	ANFIS Structure	79
4.2.2	ANFIS Learning.....	81
4.3	Information Gain.....	82
4.4	The Proposed IG –ANFIS Approach	83
4.5	The Experimental Results	84
4.6	Summary and Discussion.....	91
Chapter Five: Iterative Weighted k-NN for Constructing Missing Feature Values in Wisconsin Breast Cancer Dataset.....		92
5.1	Overview.....	92
5.2	Missing Feature Values.....	92
5.3	The Proposed Method.....	95
5.4	The Experimental Results	98

5.5	Summary and Discussion.....	100
Chapter Six: Diagnosing Breast Cancer Based on Feature Selection and Naïve Bayes..... 102		
6.1	Overview.....	102
6.2	Feature Selection Techniques	103
6.3	Feature Selection Techniques used in this Chapter.....	104
6.4	The Experiment Methodology	105
6.5	The Experimental Results	107
6.6	Summary and Discussion.....	113
Chapter Seven: Fusion of Heterogeneous Classifiers for Breast Cancer Diagnosis 114		
7.1	Overview.....	114
7.2	Multi-Classification Approach.....	115
7.2.1	Classifier Selection	115
7.2.2	Fusion Classifier	115
7.3	Classifiers Combination Strategies	116
7.4	Experimental Methodology.....	117
7.5	Experimental Results	118
7.6	Summary and Discussion.....	121
Chapter Eight: Discussion and Future Work..... 122		
References 130		

List of Figures

Figure 1: Medical Doctors per 1000 population in selected countries in Organization for Economic Cooperation and Development (OECD) Countries, 2009.	6
Figure 2: Number of MRI units per one million populations in selected countries in Organization for Economic Cooperation and Development (OECD) Countries, 2003.	7
Figure 3: Updated eHealth Architecture Including the proposed integrated intelligent system [12].....	10
Figure 4: General approach for building a classification model.....	20
Figure 5: Example of k -NN [16].....	22
Figure 6: k -NN characteristics in regards to some learning features.	26
Figure 7: Human neuron [33]	28
Figure 8: Artificial Neuron	29
Figure 9: Simplified neuron operation.....	29
Figure 10: ANN architecture	30
Figure 11: ANN characteristics in regards to some learning features.	30
Figure 12: Simple Decision Tree.....	31
Figure 13: Decision Tree characteristics in regards to some learning features.....	33
Figure 14: Bayesian classifier characteristics in regards to some learning features.	35
Figure 15: The Wrapper approach for features subset selection [65]	50
Figure 16: The filter approach [56].....	51
Figure 17: Research Method Overview	69
Figure 18: (a) Fuzzy Reasoning (b) Equivalent ANFIS Structure [89].	82
Figure 19: The general structure for the proposed approach	84
Figure 20: Information Gain Ranking on WBC	86
Figure 21: Sugeno Fuzzy Inference System with four features input and single output	87
Figure 22: Input Membership Function for the feature “Uniformity of Cell Size”	88
Figure 23: The structure for the proposed approach (IG-ANFIS)	89
Figure 24: ANFIS Structure on MATLAB.....	89
Figure 25: Comparison of classification accuracy between IG-ANFIS and some previous work	90
Figure 26: The Flowchart for the proposed method (Constructing Missing Features Values)	97
Figure 27: A comparison of classification accuracy for the proposed method through Euclidean/ k -NN.....	99

Figure 28: A comparison of classification accuracy for the proposed method through Minkoski/ k -NN	100
Figure 29: Hybrid method of feature selection technique and a learning algorithm.....	106
Figure 30: Attributes selection methods with Naïve Bayes	108
Figure 31: Results for attributes selection methods with k -NN	110
Figure 32: Results for attributes selection methods with Decision Tree	112
Figure 33: Hybrid method of feature selection technique and a learning algorithm.....	113
Figure 34: Single Classifier on three datasets WBC, WDBC, and WPBC.....	119
Figure 35: Two Classifiers on three datasets WBC, WDBC, and WPBC.	120
Figure 36: The Fusion of three classifiers on three datasets WBC, WDBC, and WPBC. ...	121
Figure 37: Results for attributes selection methods with Naïve Bayes on three databases (Thyroid, Hepatitis, and Breast Cancer)	127
Figure 38: Results for Attributes Selection Methods with k -NN on three databases (Thyroid, Hepatitis, and Breast Cancer).....	128
Figure 39: Results for Attributes Selection Methods with Decision Tree on three databases (Thyroid, Hepatitis, and Breast Cancer)	129

List of Tables

Table 1: The confusion matrix for classifier $c(x)$ on matrix X that contains 160 records. ...	21
Table 2: Examples, advantages, and disadvantages of wrapper feature selection [63].....	50
Table 3: Examples, advantages, and disadvantages of filter feature selection [63].....	52
Table 4: Examples, advantages, and disadvantages of embedded feature selection [63]	53
Table 5: Extract of data to demonstrate Expectation Maximization [83]	62
Table 6: The calculations of mean, variance, and covariance for the features depression, age, height, and weight.....	63
Table 7: The final data set after performing Expectation Maximization method.	64
Table 8: Selection of research paradigms and research methods [85]	67
Table 9: Sample of Wisconsin Breast Cancer Diagnosis dataset.....	71
Table 10: Information Gain Ranking Using WEKA on WBC.....	85
Table 11: Comparison of classification accuracy between IG-ANFIS and some previous work	90
Table 12: Results for Attributes Selection Methods with Naïve Bayes.	107
Table 13: Results for Attributes Selection Methods with k -NN	109
Table 14: Results for Attributes Selection Methods with Decision Tree.....	111
Table 15: Statistics of Breast Cancer Datasets.....	118
Table 16: Results for attributes selection methods with Naïve Bayes on three databases (Thyroid, Hepatitis, and Breast Cancer)	126
Table 17: Results for Attributes Selection Methods with k -NN on three databases (Thyroid, Hepatitis, and Breast Cancer).....	128
Table 18: Results for Attributes Selection Methods with Decision Tree on three databases (Thyroid, Hepatitis, and Breast Cancer)	129

Acronyms

ADALINE: Adaptive linear Element

ANFIS: Fuzzy Inference System

ANN: Artificial Neural Network

CAD: Computer Aided Diagnosis

CART: Classification and Regression Tree

CES: Consistency Based Subset Evaluation

CFS: correlation based feature selection

DM: Data Mining

eHealth: Electronic Health

EHR: Electronic Health Record

ERR: Error Rate

FIS: Fuzzy Inference System

GA: Genetic Algorithm

HIS: Hybrid Intelligent System

IG: Information Gain

IGANFIS: Information Gain and Adaptive Neuro-Fuzzy Inference System

k -NN: k Nearest Neighbors

LSE: Least Square Estimate

MAR: Missing At Random

MCAR: Missing Completely At Random

ML: Machine learning

MNAR: Missing Not At Random

NEHTA: National Electronic Health Transition Authority

OECD: Organization for Economic Cooperation and Development

PCA: Principle Components Analysis

R: Relief

RT: Random Tree

SBFS: Sequential Backward Floating Search

SFFS: Sequential Forward Floating Search

SFS: Sequential Forward Search

SU: Symmetrical Uncertainty

UCI Machine Learning Repository: University of California Irvine Machine Learning Repository

WBC: Wisconsin Breast Cancer Dataset

WEKA: Waikato Environment for Knowledge Analysis