ENSEMBLE FILTERS WITH HARMONIZE ALGORITHM FOR OPTIMAL SOLUTIONS IN MEDICAL DATASETS

TENGKU MAZLIN BINTI TENGKU AB HAMID

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Philosophy

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

APRIL 2021

DEDICATION

For Him, my utmost gratitude for all the miracles and strength along this journey.

For my beloved family, thank you for all the love, patience, and support.

For my fellow friends, thank you for all the friendships, spirits, and memories.

"***What comes easy won't last, what lasts won't come easy.***"

Syukur, Alhamdulillah.

ACKNOWLEDGEMENT

First and foremost, the journey of preparing this thesis was contributed along with several person, lecturers, and researchers that I have gratefully met throughout my master's life.

I specially dedicate my sincere appreciation to my main supervisor, Associate Professor Dr. Roselina Sallehuddin, for her guidance, advices, encouragement, sharing of knowledge and motivation through all my research studies. I am truly thankful to my co-supervisor, Dr. Zuriahati Mohd Yunos for her guidance, concerns, advices, and friendship along this journey. Not to forget Dr. Aida Ali, thank you for your support and sharing of knowledge in this journey. Without their assistance, patience and interest, this thesis would not have been completed as it is today.

A special thanks to Research Management Centre (RMC) of Universiti Teknologi Malaysia (UTM) for their assistance. Not to forget, many thanks to the Fundamental Research Grant Scheme (FRGS) for funding this research project under vote number R.J130000.7828.4F989 (PYL/2019/01982) and R.J130000.7851.5F154 (FRGS/1/2019/ICT02/UTM/02/13).

The highest appreciation goes to the most special person in my life for trusting and providing me support financially and spiritually, all my family members for the unconditional love, prayers, and encouragement. I would also extend my gratitude to my fellow postgraduate friends, lab mates, who have directly or indirectly lent me a help at various occasions with encouraging spirits and friendships upon completing this research. Their useful views and tips have contributed towards my understanding and thoughts indeed.

Last but not least, I would like to thank everyone who has made this journey a meaningful experience. I am truly grateful. Thank you so much.

ABSTRACT

Explosive increases of features in high dimensional datasets remains a challenge for data analysis in various research fields, especially the medical diagnosis sector, as it may affects the treatment received by the patients. Besides data dimensionality, classifiers such as Support Vector Machine (SVM) still lacks consistency in achieving an optimal performance due to improper kernel parameter settings. Commonly, the filter algorithm is frequently used for selecting relevant features due to its simple ranking strategies. However, most independent filter algorithms do not consider the intercorrelation between features, where a less dependent feature is the leading cause of why some features render irrelevant. Consequently, an imbalance number of features that could degrade the classification accuracy was produced. This problem can be alleviated using ensemble feature selection approach to identify the appropriate number of features by considering features dependency. In this study, an ensemble filters feature selection with harmonize classification algorithm has been proposed. The ensemble filters using Information Gain, Gain Ratio, Chi-squared and Relief-F are utilized with occurrence rate evaluation to identify the initial top-ranked features relevant for classification. A harmonize classification method is implemented using Particle Swarm Optimization (PSO) and SVM to synchronously determine the optimum kernel parameters and significant features as the optimal solution. The proposed method is evaluated on four medical datasets with different sizes in terms of accuracy, sensitivity, specificity, and Area under the Curve (AUC). Experimental results showed that the accuracy of the proposed method successfully increases significantly in each dataset by 96.15%, 95.41%, 96.62% and 96.50% with an optimal solution than conventional SVM. Via 10-fold cross-validation, the proposed method also signifies better classification performance compared to other existing methods. Therefore, the proposed method applies to handle high dimensional medical datasets for accurate disease prediction.

ABSTRAK

Peningkatan ciri dalam set data berdimensi tinggi kekal sebagai cabaran terhadap analisis data dalam pelbagai bidang kajian terutamanya sektor diagnosis perubatan kerana ia boleh menjejaskan rawatan yang diterima pesakit. Selain dimensi data, pengelas seperti Mesin Sokongan Vektor (SVM) masih kurang tekal dalam mencapai prestasi yang optimum akibat ketidaksesuaian penggunaan parameter kernel. Kebiasaannya, algoritma tapisan lebih kerap digunakan untuk mengenalpasti ciri-ciri relevan kerana strategi peringkat yang mudah. Namun, kebanyakan algoritma tapisan tunggal tidak dapat mengambil kira interaksi antara ciri, dimana ciri yang kurang kebergantungan ialah punca utama sesuatu ciri menjadi tidak relevan. Akibatnya, ketidakseimbangan jumlah ciri yang boleh merendahkan ketepatan pengelas dihasilkan. Masalah ini boleh diatasi menggunakan pendekatan pemilihan ciri gabungan untuk memilih jumlah ciri yang optima dengan mengambil kira kebergantungan ciri. Dalam kajian ini, satu gabungan pemilihan ciri tapisan dengan algoritma pengelasan harmoni telah dicadangkan. Gabungan tapisan menggunakan Dapatan Maklumat, Nisbah Dapatan, Persegi Chi dan Lepasan-F digunakan bersama pengiraan kadar kekerapan untuk mengenalpasti ciri awal berperingkat tinggi yang relevan untuk pengelasan. Kaedah pengelasan harmoni diterapkan menggunakan Pengoptimuman Kerumunan Zarah (PSO) dan SVM untuk mengenalpasti parameter kernel dan ciri relevan yang optimum secara serentak sebagai solusi optimal. Keberkesanan kaedah cadangan telah dinilai menggunakan empat set data perubatan yang berlainan saiz dari segi ketepatan, kepekaan, kekhususan dan kawasan dibawah keluk (AUC). Hasil kajian mendapati ketepatan kaedah cadangan berjaya meningkat kepada 96.15%, 95.41%, 96.62% dan 96.50% dengan solusi optimal oleh setiap set data berbanding SVM. Melalui keesahan bersilang 10 lipatan, kaedah cadangan juga menandakan prestasi pengelasan yang lebih baik berbanding kaedah sedia ada. Oleh itu, kaedah cadangan ini dapat digunakan dalam mengendalikan set data perubatan berdimensi tinggi untuk diagnosis penyakit yang lebih tepat.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

ACO	-	Ant Colony Optimization
ARFF	-	Attribute Relation File Format
AUC	-	Area Under the Curve
ANN	-	Artificial Neural Network
BMR	-	Boundary Margin Relief-F
CCD	-	Centre Composite Design
CFS	-	Correlation Feature Selection
CMWOA	-	Chaotic Multi-Swarm Whale Optimization Algorithm
CS	-	Chi-squared
DT	-	Decision Trees
FN	-	False Negative
FP	-	False Positive
FS	-	Fisher Score
GA	-	Genetic Algorithm
GR	-	Gain Ratio
GSA	-	Gravitational Search Algorithm
IG	-	Information Gain
KNN	-	K-Nearest Neighbours
MI	-	Mutual Information
NB	-	Naïve Bayes
PSO	-	Particle Swarm Optimization
RBF	-	Radial Basis Function
RF	-	Relief-F
SU	-	Symmetrical Uncertainty
SVM	-	Support Vector Machine
TN	-	True Negative
TP	-	True Positive
UCI	-	UC Irvine Machine Learning Repository
UTM	-	Universiti Teknologi Malaysia
WEKA	-	Waikato Environment for Knowledge Analysis

LIST OF SYMBOLS

C	-	Cost Penalty Parameter
C_1	-	Cognitive Learning Factor
C_2	-	Social Learning Factor
f	-	Frequency of Occurrence Rate
S	-	Population Size
t	-	Threshold Value
X_i	-	Original Dataset Features
X'_i	-	Top Ranked Features Output
X'_C	-	Ensemble Features Output
X_O	-	Optimum Significant Features Output
y	-	Kernel Function Parameter

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Research Overview

Medical data analysis plays a significant role to diagnose various diseases and abnormality in different parts of the body such as breast cancer, blood cancer, lymphoma cancer, skin cancer, brain cancer, hearing disabilities and etc (Chugh, 2021; Saba, 2020; Gupta and Garg, 2020). In recent times, machine learning has been widely adopted in medical sector to revolutionize the clinical decision making due to its capabilities to discover the hidden patterns of massive medical data as the supportive methods for common biopsies. As examples, cancer is the top leading cause of tumour related deaths among people in the world including Malaysian (Cancer Research Malaysia, 2021). Though, the survival rates can be improved if earlier diagnosis is conducted for early detection. Moreover, several clinical reports stated that the common imaging tests such as computerized tomography (CT) scan, magnetic resonance imaging (MRI), positron emission tomography (PET) scan, mammography, ultrasound, and X-ray are sometimes lack of high diagnostic capability and painful procedures (Adane et al., 2019 and Oskouei et al., 2017). Through machine learning, the human errors made by medical experts during diagnosis can be reduced by extracting and processing the information in medical data precisely in less required time.

One of the prevalent machine learning models that have been widely applied in medical data analysis is classification. Due to explosive increase of medical data, the amount of disease information has become accumulated into high dimensional data which resulting a complexity issues in medical diagnosis (Gupta and Garg, 2020; Garba and Harande, 2018). Such massive amount of data could not be processed efficiently for an accurate prediction. Furthermore, the presence of irrelevant features and redundant information in medical data are not considered properly in most studies.

Consequently, the classification accuracy may be degraded by the existence of irrelevant features and indirectly increase the computational time for diagnosis (Ghorbani and Ghousi, 2019). Thus, the extraction of useful information from medical data is highly required for improving diagnosis and treatments.

Apart from data dimensionality, the performance of classifier still can be influenced by the settings of kernel parameters in the training process (Raja and Pandian; 2020; Zhong et al., 2017; Sallehuddin et al., 2016). This shows that a proper medical data analysis performance particularly relied on the quality of input data and the parameters of classifiers (Oskouei et al., 2017; Omar et al., 2012). Therefore, this research attempts to improve the classification performance by utilizing feature selection prior to classification to first identify the irrelevant or redundant features and then determine the optimum significant features and classification parameters for optimal solution. Support Vector Machine (SVM) is employed as classifier based on its robust performance in avoiding local minima and overfitting solutions.

1.2 Problem Background

Generally, various features are used in medical datasets to represent various disease prediction and medical diagnosis through classification. Data pre-processing such as feature selection is a significant process to explore the medical information since the performance of classifier influenced by the quality of medical data known as the training samples (Omar et al., 2013; Ubaidillah et al., 2013). However, such training samples tend to be ambiguous when an explosive number of input features expands. In addition, certain features may consist of irrelevant and redundant information that increases the dimensionality of medical data. An increased of dimensionality also resulting a complexity in processing the algorithm (Ali et al., 2019; Miao and Niu, 2016). As a result, the memory space and computational time are highly consumed in processing the algorithm which indirectly degrade the accuracy of the classifier. This situation has resulting challenges in diagnosing disease and interpreting data due to the inconsistency and confusing data patterns (Singh et al., 2016; Taylor and Kim, 2011; Wang and Ma, 2009).

Several reports stated that the irrelevant and redundant features in high dimensional data are required to be eliminated to address the dimensionality issues (Zhong et al., 2017; Ghaemi et al., 2016; Singh et al., 2016). In contrast, features with highest significance need to be identified in order to improve the classification performance. Thus, an improved classification model with intelligent feature selection is required for handling and exploring the high dimensional medical data. The classification model should perceive the ability to perform an accurate and computationally effective diagnosis with significant number of input features. Since large amount of data can negatively affect the classification process, it is observed that a reduced set of features is sufficient to improve the accuracy of prediction. This suggested that not all input features are relevant to be include in the training task as the classification may result a low performance when massive input features increase in the classifier (Prasad et al., 2018; Ghaemi et al., 2016). For such reasons, feature selection approach is important to pre-process the data before classification task and must be considered to produce an effective medical data analysis.

Feature selection refers to the process of selecting subset of features from a set of original features to represent the data. It is an important process in reducing data with high dimensionality by eliminating the redundant and irrelevant features that may misguide the classification performance. Feature selection can be categorized into filter, wrapper, and embedded algorithms (Zhong et al., 2017; Miao and Niu, 2016; Canedo et al., 2014; Guyon and Elisseeff, 2003). Based on literature review conducted, the filter algorithm has outperformed the wrapper and embedded algorithms in terms of less computational complexity (Lyu et al, 2017; Chandrashekar and Sahin, 2014; Hira and Gillies, 2014; Shardlow, 2011). The filter algorithm perceives the ability of improving the classification accuracy by evaluating the significance value of each input features using specific statistical measure or ranking evaluation. This made the filter algorithm less complex and computationally faster since it does not involve any classifier algorithm which is suitable for handling high dimensional data with explosive number of features (Bommert et al., 2020; Zhang et al., 2019; Hancer et al., 2018). Listed are examples of the common filter algorithms recommended for medical data analysis such as Information Gain, Gain Ratio, Chi-squared and Relief-F (Bommert et al., 2020; Zhang et al., 2019; Urbanowicz et al., 2018; Fahrudin et al., 2016).

However, independent filter algorithm can be afflicted by several limitations. The major disadvantage of independent filter algorithm is the limited correlation between features (Chandrashekar and Sahin, 2014; Omar et al., 2014; Miao and Niu, 2016; Bommert et al., 2020). This is because most independent filter algorithms only focused on evaluating the intrinsic characteristics of features and neglecting the interactions between each input features. As a result, the intercorrelation between features and features dependency is not considered in selecting features, but it produced less correlated features (Bommert et al., 2020; Hira and Gillies, 2015; Nancy and Balamurugan, 2013). Moreover, imbalanced number of significant features are produced which causing the classifier to produce inaccurate prediction. For this reason, this research is motivated to utilize an assemble of multi filters algorithm for feature selection to effectively eliminate any irrelevant and redundant input features prior to classification.

Apart from the imbalance number of features, the performance of classifier such as SVM can also be influenced by the settings of kernel parameters values in the classification tasks (Wang and Chen, 2020; Huang et al., 2018; Yan and Jia, 2018). The commonly used kernel in SVM classifier is known as Radial Basis Function (RBF), where it requires two kernel parameters named kernel function parameter (y) and soft margin constant or the penalty factor (C) in order to perform the training task (Wang and Chen, 2020; Huang et al., 2018; Hsu et al., 2016). The classification accuracy of SVM can dramatically decrease if the selection of these parameters is not properly selected. At the same time, the selection of significant features can also be affected due to improper values of C and y. Hence, it is necessary to optimize the selection of kernel parameters for accurate and optimal SVM classification.

Various optimization algorithms have been employed to provide the optimum searching solution in determining the best kernel parameters for SVM classification model. According to recent studies, Particle Swarm Optimization (PSO) is one of the most recommended searching methods for optimization due to its easy implementation and adaptability to integrate with any classifier algorithms (Ghorbani and Ghousi, 2019; Raj et al., 2018; Zhang et al., 2018). Due to its less parameter usage and faster convergence rate, PSO can perceive better optimization ability effectively compared to other algorithms such as Genetic Algorithm (GA) and Ant Colony Optimization (ACO) which consumed much higher memory space due to high parameter usage and computational complexity (Moslehi and Haeri, 2019; Sakri et al., 2018; Neha and Vashishtha, 2016). Based on this advantage, PSO is employed synchronously with SVM classification for optimizing the kernel parameters of SVM in order to obtain the optimal solution.

At the same time, the process of optimizing SVM kernel parameters may also influenced the selection of significant features (Zhang et al., 2019; Huang et al., 2018; Neha and Vashishtha, 2016; Huang and Dun., 2008). Recently, several studies reported that solution for synchronous optimization on both processes are highly suggested to determine the optimum number of significant features and kernel parameters simultaneously without affecting the classification accuracy. Due to the imbalance selection of features, poor settings in kernel parameters and the incremented of computational complexity, the requirement for harmonize classification has becomes essential (Wang and Chen, 2020; Zeng et al., 2018; Tarle et al., 2016). For this reason, this research is motivated to implement a harmonize classification method using PSO and SVM to optimize the selection of significant features and kernel parameters synchronously without minimizing the accuracy so that an optimal solution of high dimensional medical data classification can be achieved.

Based on aforementioned problems and issues, several research gaps have been identified. Firstly, most independent filter algorithm only focused on evaluating the intrinsic characteristics of features and neglecting features interactions (Bommert et al., 2020; Ali et al., 2019; Zhong et al., 2017 & Singh et al., 2016). This indicates that independent filter algorithm still lack consideration on features dependency. In consequence, imbalance number of selected features that contribute to inaccurate classifier prediction accuracy are produced, which made it difficult to observe features that truly significant for classification. Secondly, the tuning of SVM parameters using grid search method required high parameters range which could led to computationally prohibitive and sometimes infeasible (Wang & Chen, 2020; Huang et al., 2018; Srisukkham et al., 2017). Hence, an optimal classification accuracy is impossible to be achieved when the optimization and classification processes are performed separately. For this reason, this research is motivated to propose an ensemble filters feature selection using Information Gain (IG), Gain Ratio (GR), Chi-squared (CS) and Relief-F (RF) to effectively eliminate irrelevant features prior to classification without neglecting features dependency by considering the features occurrence and implement a harmonize classification method using PSO and SVM to synchronously optimize the selection of significant features and kernel parameters without degrading the accuracy based on Centre Composite Design (CCD) search method for optimal solution.

In brief, the selection of optimum significant features from high dimensional data and a proper setting of SVM parameters relatively contribute an impact towards the classification accuracy performance. It is highly important to control the quantity of input features for producing an accurate prediction and computationally low intensive classification model (Wang et al., 2019; Raj et al., 2016; Zhang et al., 2013). Besides, with an optimal number of features and kernel parameters, the classification model such as SVM can be generalized easily (Moslehi and Haeri, 2019; Huang et al., 2018; Aladeemy et al., 2017). Thus, the utilization of ensemble filters feature selection is highly necessary to identify the top significant features candidates for enhancing the efficiency of synchronous optimization as the optimal solution. Overall, the proposed method aims to improve the classification accuracy of high dimensional medical data by effectively determine the optimal solution of SVM parameters and optimum number of significant features appropriate for classification without decreasing the accuracy.

1.3 Problem Statement

In machine learning, SVM classifier is one of the best predictive models that have been widely applied in medical data analysis due to its robust performances. However, an explosive increase of information and various input features has resulting high dimensionality issues with the existence of redundant and irrelevant features which indirectly diminish the classification performance (Zhang et al., 2019; Prasad et al., 2018). Regarding this, an appropriate diagnosis prediction has become challenging since the classification accuracy is highly depends on the quality of the medical data. Thus, a reliable data pre-processing technique such as feature selection is required to improve the classification accuracy performance since it perceives the ability in handling features ambiguity and relevancy by evaluating the significance value of each input features before entering the classification process.

An independent filter feature selection often selected an unbalance number of features which made it difficult to observe the features which are truly significant for classification (Wang and Chen, 2020; Yan and Jia, 2018; Huang et al., 2018). Due to the unbalance selected features, SVM consequently failed to select a proper settings of kernel parameters and tends to produce a low classification performance when the data dimensionality increases (Zhang et al., 2019; Raj et al., 2016). This observed that the unbalanced number of selected features and improper selection of SVM parameters may consequently degrade the accuracy of classification performance (Prasad et al., 2018; Han and Bian, 2018). Hence, an improved classification model that could dynamically produce the highest classification accuracy with optimal solution of classification parameters and optimal significant features is highly demanded.

In addition, the process of feature selection and kernel parameters settings are dependent, in which an optimal SVM classification accuracy are most likely impossible to be achieved when both processes are performed separately (Wang and Chen, 2020; Huang et al., 2018). This problem can be alleviated by implementing optimization method in searching for optimal solution. According to studies, PSO is the most recommended searching method for optimization due to its capability for parallel processing (Wang & Chen, 2020; Huang et al., 2018; Srisukkham et al., 2017). Since the value of SVM parameters may influence the selection of significant features, it is necessary to determine the best SVM parameters and optimal number of significant features simultaneously. Thus, an improved SVM classification model with reliable feature selection and parameter optimization method must be developed to produce an accurate medical diagnosis prediction without degrading the classification accuracy. The following hypothesis were derived to support the problem statement:

"The accuracy of SVM classification model can be improved effectively by utilizing ensemble filters feature selection with occurrence rate evaluation and harmonize classification of PSO and SVM for optimal solution in high dimensional medical datasets."

1.4 Research Questions

The research questions to support the hypothesis statement are as follows:

(a) How does the ensemble filters feature selection identify the top ranked features and eliminate the irrelevant features from medical datasets?

(b) How does the harmonize classification algorithm of PSO and SVM optimize the SVM parameters and selected features synchronously without affecting the accuracy?

(c) Does the proposed method successfully improve the accuracy using optimum SVM parameters and significant features as the optimal solution of medical datasets?

1.5 Research Aim

This research aims to improve the classification accuracy with optimal solution in high dimensional medical datasets using ensemble filters feature selection with harmonize classification algorithm. Ensemble filters feature selection using IG, GR, CS and RF is developed with occurrence rate evaluation to identify the initial top ranked features significant for classification. Then, harmonize classification using PSO and SVM algorithm is employed as the optimal solution of medical datasets to synchronously determine the optimum classification parameters and significant features without degrading the accuracy.

1.6 Research Objectives

The research objectives are presented as follows:

(a) To propose ensemble filters feature selection using IG, GR, CS and RF with occurrence rate evaluation in order to identify the initial top ranked features from medical datasets relevant for classification.

(b) To implement a harmonize classification algorithm using PSO and SVM to synchronously determine the optimum SVM parameters and optimum significant features from medical datasets with the highest training fitness.

(c) To improve the classification accuracy of medical datasets using ensemble filters feature selection with harmonize classification algorithm of PSO and SVM in searching for optimal solution.

1.7 Research Scopes

The scopes of this research are presented as follows:

(a) This research is analysed on four standard medical datasets with different dimensionality such as Breast Cancer dataset, Wisconsin Diagnostic Breast Cancer (WDBC) dataset, Lymphography dataset and Audiology dataset retrieved from UCI Machine Learning Repository at *https://archive.ics.uci.edu/ml/datasets.php*.

(b) This research utilized four filter algorithms such as Information Gain (IG), Gain Ratio (GR), Chi-squared (CS) and Relief-F (RF) as ensemble feature selection to identify the initial top ranked features of medical datasets.

(c) This research implemented PSO and SVM algorithm for harmonize classification to determine the optimal solution of medical datasets.

(d) This research is focused on employing SVM classifier to train and classify the optimum features of medical datasets into respective medical diagnosis.

1.8 Research Significance

Generally, the research is conducted to discover solutions to a certain issue in medical data analysis. This research proposed a machine learning approach using ensemble filters feature selection with harmonize classification of PSO and SVM to improve the classification accuracy in high dimensional medical datasets. The privilege of adopting machine learning in medical data analysis will contributes to medical center, medical institution and hospitals to significantly improve the reliability of high dimensional medical datasets in diagnosing diseases. The utilization of ensemble filters feature selection will assists the medical experts in identifying the top ranked relevant information out of the existing information. The significance of this research is to observe whether is it possible that medical data provides an important indicator to determine certain diseases as well as improving the prediction accuracy. Most research are focusing on classifying the medical data without emphasizing about the optimum number of significant information and improper classification parameters that must be address for establishing a reliable classifier with useful information.

Another significance of this research is to discover how successful a classifier with ensemble filters feature selection and harmonize classification based on the improvement of classification accuracy in obtaining the optimal solution from the high dimensional medical datasets. Besides that, the percentage of dimensionality reduction and classification performance are evaluated to illustrate the unseen interrelationship between optimal significant features in high dimensional medical datasets and accuracy performance for understandable clarification. This research is highly beneficial in healthcare industry especially the Malaysia’s Ministry of Health, or Cancer Research Malaysia, where the prediction of disease probability such as cancer, coronavirus disease 2019 (COVID-19) and other diseases are highly concerned. Based on the human perspectives, a reliable medical data which consists of patients' health information can be referred by both patients and families so that any possibilities regarding the disease progress can be prepared properly. Therefore, the proposed method using ensemble filters feature selection with harmonize classification algorithm of PSO and SVM is essential in determining the optimal solution of medical datasets and possible in providing alternatives for disease diagnosis in Malaysia.

1.9 Organization of Thesis

The thesis is organized into six chapters. Chapter 1 presents a brief explanation on research overview, problem background, problem statements, research questions, research objectives, research scope and research significance. Chapter 2 presents literature reviews on related machine learning algorithms where the reviews of current techniques and limitations on medical data analysis are described. Based on the literature gathered, the solution to address the problems is presented. Chapter 3 explained the research methodology where all steps and processes involved in each phase is presented. Chapter 4 presents the development of ensemble filters feature selection. Chapter 5 presents the development of harmonize classification of PSO and SVM. The evaluation and validation of classification performance towards experimental datasets are presented in this chapter. Lastly, the research findings and recommendations for future works are discussed and concluded in Chapter 6.

REFERENCES

Adane, K., Gizachew, M. and Kendie, S. (2019) 'The role of medical data in efficient patient care delivery: a review', *Risk Management and Healthcare Policy,* 12, pp. 67–73.

Ahmad, A. (2009) *Data Transformation For Decision Tree Ensembles*. PhD Thesis, University of Manchester, UK.

Ahmed, A.A., Bakar, A.A. and Hamdan, A.R. (2009) Dynamic Data Discretization Technique Based on Frequency and K-Nearest Neighbour Algorithm. *2nd Conference on Data Mining and Optimization*. 27-28 October. Selangor, Malaysia: IEEE, pp. 10-14.

Ahn, S., Kim, G. and Kim, M. (2006) 'A note on applications of support vector machine', *Mathematics Subject Classification*.

Aladeemy, M., Tutun, S. and Khasawneh, M.T. (2017) 'A new hybrid approach for feature selection and support vector machine model selection based on self-adaptive cohort intelligence', *Expert Systems with Applications*, 88, pp. 118–131.

Ali, H., Salleh, M.N.M., Saedudin, R., Hussain, K. and Mushtaq, M.F. (2019) 'Imbalance class problems in data mining: a review', *Indonesian Journal of Electrical Engineering and Computer Science*, 14(3), pp. 1560–1571.

Alirezanejad, M., Enayatifar, R., Motameni, H. and Nematzadeh, H. (2019) 'Heuristic filter feature selection methods for medical datasets', *Genomics*, 112(2), pp. 1173–1181.

Apolloni, J., Leguizamón, G. and Alba, E. (2016) 'Two hybrid wrapper-filter feature selection algorithms applied to high-dimensional microarray experiments', *Applied Soft Computing Journal*, 38, pp. 922–932.

Assarzadeh, Z. and Nilchi, A.R.N. (2015) 'Chaotic particle swarm optimization with mutation for classification', *Journal of Medical Signals and Sensors*, 5(1), pp. 12–20.

Bakar, A.A., Othman, Z.A. and Shuib, N.L.M. (2009) Building A New Taxonomy for Data Discretization Techniques. *2nd Conference on Data Mining and Optimization*. 27-28 October. Selangor, Malaysia: IEEE, pp. 138-146.

Battineni, G., Chintalapudi, N. and Amenta, F. (2019) ‘Machine learning in medicine: Performance calculation of dementia prediction by support vector machines (SVM)’, *Informatics in Medicine Unlocked,* 16, 100200.

Blum, A.L. and Langley, P. (1997) ‘Artificial intelligence selection of relevant features and examples in machine’, *Artificial Intelligence*, 97, pp. 245–271.

Bommert, A., Sun, X., Bischl, B., Rahnenfuhrer, J. and Lang, M. (2020) ‘Benchmark for filter methods for feature selection in high-dimensional classification data’, *Computational Statistics and Data Analysis*, 143, 106839.

Cancer Research Malaysia. (2021) *Cancers* [online]. Available at: https://www.cancerresearch.my/our-work/cancers/ (Accessed: 1 March 2021).

Canedo, V.B., Marono, N.S. and Betanzos, A.A. (2012) ‘An ensemble of filters and classifiers for microarray data classification’, *Pattern Recognition*, 45(1), pp. 531–539.

Canedo, V.B., Marono, N.S., Betanzos, A.A., Benitez, J.M. and Herrera, F. (2014) ‘A review of microarray datasets and applied feature selection methods’, *Information Sciences*, 282, pp. 111–135.

Chandrashekar, G. and Sahin, F. (2014) ‘A survey on feature selection methods’, Computers and Electrical Engineering. 40, 16–28.

Chugh, G., Kumar, S. and Singh, N. (2021) ‘Survey on machine learning and deep learning applications in breast cancer diagnosis’, *Cognitive Computation,* (0123456789).

Cortes, C. and Vapnik, V. (1995) ‘Support-vector networks’, *Machine Learning*, 20, pp. 273-297.

Dag, H., Sayin, K.E., Yenidogan, I., Albayrak, S. and Acar, C. (2012) Comparison Of Feature Selection Algorithms for Medical Data. *International Symposium on Innovations in Intelligent Systems and Applications*. 2-4 July. Trabzon, Turkey: IEEE, pp. 1–5.

Dai, J. and Xu, Q. (2013) ‘Attribute selection based on information gain ratio in fuzzy rough set theory with application to tumor classification’, *Applied Soft Computing*, 13, pp. 211–221.

Dankolo, M.N., Radzi, N.H.M., Sallehuddin, R. and Mustaffa, N.H. (2017) A Study of Metaheuristic Algorithms for High Dimensional Feature Selection on Microarray Data. *Proceedings of the 13th IMT-GT International Conference* *on Mathematics, Statistics and Their Applications (ICMSA2017)*. 4-7 December. Kedah, Malaysia: AIP Conference Proceedings, 1905(1), 040010.

Das, A.K., Goswami, S., Chakrabarti, A. and Chakraborty, B. (2017) ‘A new hybrid feature selection approach using feature association map for supervised and unsupervised classification’, *Expert Systems with Applications*, 88, pp. 81–94.

Deepika, M., Mary G.L. and Madhu K.R. (2016) ‘A review on prediction of breast cancer using various data mining techniques’, *Research Journal of Pharmaceutical, Biological and Chemical Sciences*, 7(1), pp. 808–814.

Dongare, S.A., Ande, V.K. and Tirandasu, R.K. (2018) ‘A feature selection approach for enhancing the cardiotocography classification performance’, *International Journal of Engineering and Techniques,* 4(2), pp. 222–226.

Fahrudin, T.M., Syarif, I. and Barakbah, A.R. (2016) Feature Selection Algorithm Using Information Gain Based Clustering for Supporting the Treatment Process of Breast Cancer. *Proceeding of the International Conference on Informatics and Computing (ICIC)*. 28-29 October. Mataram, Indonesia: IEEE, pp. 6–11.

Fonti, V. (2017) 'Feature Selection using LASSO', *Research Paper in Business Analytics,* Vrije Universiteit Amsterdam.

Garba, K.D. and Harande, Y.I. (2018) ‘Significance and challenges of medical records: a systematic literature review’, *Journal of Advances in Librarianship*, 9(1), pp. 26–31.

Ghaemi, M. and Feizi-Derakhshi, M.R. (2016) ‘Feature selection using forest optimization algorithm’, *Pattern Recognition*, 60, pp. 121–129.

Ghimatgar, H., Kazemi, K., Helfroush, M.S. and Aarabi, A. (2018) ‘An improved feature selection algorithm based on graph clustering and ant colony optimization,’ *Knowledge-Based Systems*, 159, pp. 270–285.

Ghorbani, R. and Ghousi, R. (2019) ‘Predictive data mining approaches in medical diagnosis: A review of some disease prediction’, *International Journal of Data and Network Science*, 3, pp. 47–70.

Gupta, P. and Garg, S. (2020) ‘Breast cancer prediction using varying parameters of machine learning models’, *Procedia Computer Science*, 171, pp. 593–601.

Guyon, I. and Elisseeff, A. (2003) ‘An introduction to variable and feature selection’, *Journal of Machine Learning Research*, 3, pp. 1157–1182.

Hairuddin, N.L., Yusuf, L.M., Othman, M.S. and Majid, H.A. (2016) ‘Improving gender classification with feature selection in forensic anthropology’, *Jurnal Teknologi*, 12(2), pp. 57–63.

Hameed, S.S., Petinrin, O.O., Hashi, A.O. and Saeed, F. (2018) "Filter-wrapper combination and embedded feature selection for gene expression data', *International Journal of Advances in Soft Computing and its Applications*, 10(1), pp. 90–105.

Hamid, T.M.T.A., Sallehuddin, R., Yunos, Z.M. and Ali, A. (2019) ‘Ensemble based multi filters algorithm for tumor classification in high dimensional microarray dataset’, *International Journal of Advanced Trends in Computer Science and Engineering*, 8(1.6), pp. 116–123.

Han, B. and Bian, X. (2018) ‘A hybrid PSO-SVM-based model for determination of oil recovery factor in the low-permeability reservoir’, *Petroleum*, 4(1), pp. 43–49.

Han, J. and Kamber, M. (2001) *Data mining: concepts and techniqu*es. 3rd edn. USA: Morgan Kaufmann Publishers.

Han, J., Kamber, M. and Pei, J. (2012) *Classification: advanced methods*, in *Data mining: concepts and techniques*. 3rd edn. USA: Morgan Kaufmann Publishers, pp. 393–442.

Hancer, E., Xue, B. and Zhang, M. (2018) ‘Differential evolution for filter feature selection based on information theory and feature ranking’, *Knowledge-Based Systems*, 140, pp. 103–119.

Hand, J., Mannila, H. and Smyth, P. (2001) *Principles of data mining*. London: The MIT Press.

Harb, H.M. and Desuky, A.S. (2014) ‘Feature selection on classification of medical datasets based on particle swarm optimization’, *International Journal of Computer Applications*, 104(5), pp. 14–17.

Hira, Z.M. and Gillies, D.F. (2015) ‘A review of feature selection and feature extraction methods applied on microarray data,’ *Advances in Bioinformatics*, 198363.

Huang, C.L. and Dun, J.F. (2008) ‘A distributed PSO-SVM hybrid system with feature selection and parameter optimization,’ *Applied Soft Computing Journal*, 8, pp. 1381–1391.

Huang, S., Cai, N., Pacheco, P.P., Narandes, S., Wang, Y. and Xu, W. (2018) ‘Applications of support vector machine (SVM) learning in cancer genomics’, *Cancer Genomics and Proteomics*, 15, pp. 41–51.

Hulse, J.V., Khoshgoftaar, T.M., Napolitano, A. and Wald, R. (2012) ‘Threshold-based feature selection techniques for high-dimensional bioinformatics data’, *Network Modeling and Analysis in Health Informatics and Bioinformatics*, 1, pp. 47–61.

Hsu, C.W., Chang, C.C. and Lin, C.J. (2016) ‘A practical guide to support vector classification’, *Technical Report, Department of Computer Science National Taiwan University* 20(April), pp. 1-16.

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

Jayaram, M.A., Karegowda, A.G. and Manjunath, A.S. (2010) ‘Feature subset selection problem using wrapper approach in supervised learning’, *International Journal of Computer Applications*, 1(7), pp. 13–17.

Kahng, J., Kim, E.H., Kim, H.G. and Lee, W. (2015) ‘Development of a cervical cancer progress prediction tool for human papillomavirus-positive Koreans: A support vector machine-based approach’, *Journal of International Medical Research*, 43(4), pp. 518–525.

Karegowda, A.G., Jayaram, M.A. and Manjunath, A.S. (2011) ‘Feature subset selection using cascaded GA and CFS: a filter approach in supervised learning’, *International Journal of Computer Applications*, 23(2), pp. 1–10.

Karegowda, A.G., Manjunath, A.S. and Jayaram, M.A. (2010) ‘Comparative study of attribute selection using gain ratio and correlation based feature selection’, *International Journal of Information Technology and Knowledge and Knowledge Management*, 2(2), pp. 271–277.

Kennedy, J. and Eberhart, R.C. (1995) Particle Swarm Optimization. *Proceedings of International Conference on Neural Network (ICNN’95).* 27 November-1 December. Perth, Australia: IEEE, pp. 1942-1948.

Konsiantis, S.B. (2007) ‘Supervised machine learning: a review of classification techniques’, *Informatica*, 31, pp. 249–268.

Lee, I.H., Lushington, G.H. and Visvanathan, M. (2011) ‘A filter-based feature selection approach for identifying potential biomarkers for lung cancer’, *Journal of Clinical Bioinformatics*, 1(11), pp. 1–8.

Lyu, H., Wan, M., Han, J., Liu, R. and Wang, C. (2017) ‘A filter feature selection method based on the maximal information coefficient and gram-schmidt orthogonalization for biomedical data mining’, *Computers in Biology and Medicine*, 89, pp. 264–274.

Ma, S. and Huang, J. (2008) ‘Penalized feature selection and classification in bioinformatics’, *Briefings in Bioinformatics*, 9(5), pp. 392–403.

Miao, J. and Niu, L. (2016) ‘A survey on feature selection’, *Procedia Computer Science*, 91, pp. 919–926.

Mohamed, N.S., Othman, Z.A. and Bakar, A.A. (2009) A Classification of “*Graciliria changii*” Protein Sequences Using Back-Propagation Classifier. *2nd Conference on Data Mining and Optimization*. 27-28 October. Selangor, Malaysia: IEEE, pp. 94-99.

Moslehi, F. and Haeri, A. (2019) ‘An evolutionary computation-based approach for feature selection’, *Journal of Ambient Intelligence and Humanized Computing*, 11, pp. 3757–3769.

Mutalib, S., Ali, N.A., Rahman, S.A. and Mohamed, A. (2009) An Exploratory Study in Classification Methods for Patients’ Dataset. *2nd Conference on Data Mining and Optimization*. 27-28 October. Selangor, Malaysia: IEEE, pp. 86-90.

Nagpal, S., Arora, S. and Dey, S. (2017) ‘Feature selection using gravitational search algorithm for biomedical data’, *Procedia Computer Science*, 115, pp. 258–265.

Nancy, S.G. and Balamurugan, S.A. (2013) ‘A comparative study of feature selection methods for cancer classification using gene expression dataset’, *Journal of Computer Applications (JCA)*, 6(3), pp. 78-84.

Neha and Vashishtha, J. (2016) ‘Particle swarm optimization based feature selection’, *International Journal of Computer Applications*, 146(6), pp. 11–17.

Nekkaa, M. and Boughaci, D. (2012) Improving Support Vector Machine Using A Stochastic Local Search for Classification in Data Mining. *International*

Conference on Neural Information Processing (ICONIP 2012). 12-15 November. Doha, Qatar: Springer, pp. 168–175.

Omar, N. (2014) *Feature Selection for Classification of Survival Analysis in Lymphoma Cancer*. Master Thesis, Universiti Teknologi Malaysia, Skudai.

Omar, N., Jusoh, F., Othman, M.S. and Ibrahim, R. (2013) ’Review of feature selection for solving classification problems’, *Journal of Information Systems Research and Innovation*, pp. 64-70.

Omar, N., Othman, M.S., Ibrahim, R. and Jusoh, F. (2012) ‘Particle swarm optimization feature selection for classification of survival analysis in lymphoma cancer’, *International Journal of Innovative Computing*, 2(1), pp. 1–9.

Osanaiye, O., Cai, H., Choo, K.K.R., Dehghantanha, A., Xu, Z. and Dlodlo, M. (2016) ‘Ensemble-based multi-filter feature selection method for DDoS detection in cloud computing’, *EURASIP Journal on Wireless Communications and Networking*, 130, pp. 1–10.

Oskouei, R.J., Kor, N.M. and Maleki, S.A. (2017) ‘Data mining and medical world: breast cancers’ diagnosis, treatment, prognosis and challenges’, *American Journal of Cancer Research*. 7(3), pp. 610–627.

Pardo, B.S., Canedo, V.B. and Betanzos, A.A. (2019) ‘On developing an automatic threshold applied to feature selection ensembles’, *Information Fusion*, 45, pp. 227–245.

Prasad, Y., Biswas, K.K. and Hanmandlu, M. (2018) ‘A recursive PSO scheme for gene selection in microarray data’, *Applied Soft Computing*, 71, pp. 213–225.

Rahman, S.A., Bakar, A.A. and Hussein, Z.A.M. (2009) Filter-Wrapper Approach to Feature Selection Using RST-DPSO for Mining Protein Function. *2nd Conference on Data Mining and Optimization*. 27-28 October. Selangor, Malaysia: IEEE, pp.71–78.

Raj, D.M.D. and Mohanasundaram, R. (2020) ‘An efficient filter-based feature selection model to identify significant features from high-dimensional microarray data’, *Arabian Journal for Science and Engineering*, 45, pp. 2619–2630.

Raj, S., Ray, K.C. and Shankar, O. (2016) ‘Cardiac arrhythmia beat classification using DOST and PSO tuned SVM’, *Computer Methods and Programs in Biomedicine*, 136, pp. 163–177.

Raja, J.B. and Pandian, S.C. (2020) 'PSO-FCM based data mining model to predict diabetic disease', *Computer Methods and Programs in Biomedicine*, 196, 105659.

Rani, R.R. and Ramyachitra, D. (2018) ‘Microarray cancer gene feature selection using spider monkey optimization algorithm and cancer classification using SVM’, *Procedia Computer Science*, 143, pp. 108–116.

Saba, T. (2020) 'Recent advancement in cancer detection using machine learning: Systematic survey of decades, comparisons and challenges', *Journal of Infection and Public Health*, 13(9), pp. 1274–1289.

Sakri, S., Rashid, N.A. and Zain, Z.M. (2018) ‘Particle swarm optimization feature selection for breast cancer recurrence prediction’, in *IEEE Access*, 6, pp. 29637–29647.

Sallehuddin, R., Ubaidillah, S.H.S.A., Zain, A.M., Alwee, R. and Radzi, N.H.M. (2016) ‘An improvement in support vector machine classification model using grey relational analysis for cancer diagnosis’, *Jurnal Teknologi*, 78(8–2), pp. 107–119.

Salzberg, S.L. (1997) ‘On comparing classifiers: Pitfalls to avoid and a recommended approach’, *Data Mining and Knowledge Discovery*, 1, pp. 317–328.

Samant, R. and Rao, S. (2013) ‘A study on feature selection methods in medical decision support systems’, *International Journal of Engineering Research and Technology*, 2(11), pp. 615–620.

Santos, V., Datia, N. and Pato, M.P.M. (2014) ‘Ensemble feature ranking applied to medical data’, *Procedia Technology*, 17, pp. 223–230.

Selvakuberan, K., Kayathiri, D., Harini, B. and Devi, M.I. (2011) An Efficient Feature Selection Method for Classification in Health Care Systems Using Machine Learning Techniques. *3rd International Conference on Electronics Computer Technology*. 8-10 April. Kanyakumari, India: IEEE, pp. 223–226.

Shardlow, M. (2011) ‘An analysis of feature selection techniques’, *Conference Proceedings*, pp.1–7.

Singh, B., Gornet, M., Sims, H., Kisanga, E., Knight, Z. and Segars, J. (2020) ‘Severe Acute Respiratory Syndrome Coronavirus 2 (SARS-CoV-2) and its effect on gametogenesis and early pregnancy’, *American Journal of Reproductive Immunology*, 84(5), pp. 1–9.

Singh, B.K., Verma, K. and Thoke, A.S. (2016) ‘Fuzzy cluster based neural network classifier for classifying breast tumors in ultrasound images’, *Expert Systems with Applications*, 66, pp. 114–123.

Srisukkham, W., Zhang, L., Neoh, S.C., Todryk, S. and Lim, C.P. (2017) ‘Intelligent leukaemia diagnosis with bare-bones PSO based feature optimization’, *Applied Soft Computing*, 56, pp. 405–419.

Tarle, B., Tajanpure, R. and Jena, S. (2016) ‘Medical data classification using different optimization techniques: a survey’, *International Journal of Research in Engineering and Technology*, 5(5), pp. 101–108.

Taylor, S.L. and Kim, K. (2011) ‘A jackknife and voting classifier approach to feature selection and classification’, *Cancer Informatics*, 10, pp. 133–147.

Trotter, M.W.B., Buxton, B.F. and Holden, S.B. (2001) ‘Support vector machines in combinational chemistry’, *Measurement & Control*, 34, pp. 235-239.

Tuba, E., Strumberger, I., Bezdan, T., Bacanin, N. and Tuba, M. (2019) ‘Classification and feature selection method for medical datasets by brain storm optimization algorithm and support vector machine’, *Procedia Computer Science*, 162(3), pp. 307–315.

Ubaidillah, S.H.S.A., Sallehuddin, R. and Ali, N.A. (2013) ‘Cancer detection using aritifical neural network and support vector machine: a comparative study’, *Jurnal Teknologi (Sciences and Engineering)*, 65(1), pp. 73–81.

Urbanowicz, R.J., Olson, R.S., Schmitt, P., Meeker, M. and Moore, J.H. (2018) ‘Benchmarking relief-based feature selection methods for bioinformatics data mining’, *Journal of Biomedical Informatics*, 85, pp. 168–188.

Wang, M. and Chen, H. (2020) ‘Chaotic multi-swarm whale optimizer boosted support vector machine for medical diagnosis’, *Applied Soft Computing Journal*, 88, 105946.

Wang, X., Guan, S., Hua, L., Wang, B. and He, X. (2019) ‘Classification of spot-welded joint strength using ultrasonic signal time-frequency features and PSO-SVM method’, *Ultrasonics*, 91, pp. 161–169.

Wang, Y. and Ma, L. (2009) 'FF-based feature selection for improved classification of medical data', *WSEAS Transactions on Computer,* 8(2), pp. 396–405.

Xi, M., Sun, J., Liu, L., Fan, F. and Wu, X. (2016) 'Cancer feature selection and classification using a binary quantum-behaved particle swarm optimization and support vector machine', *Computational and Mathematical Methods in Medicine,* 2016, pp. 1–9.

Yan, X. and Jia, M. (2018) 'A novel optimized SVM classification algorithm with multi-domain feature and its application to fault diagnosis of rolling bearing', *Neurocomputing*, 313, pp. 47–64.

Zeng, N., Qiu, H., Wang, Z., Liu, W., Zhang, H. and Li, Y. (2018) 'A new switching-delayed-PSO-based optimized SVM algorithm for diagnosis of Alzheimer's disease', *Neurocomputing*, 320, pp. 195–202.

Zhang, J., Xiong, Y. and Min, S. (2019) 'A new hybrid filter/wrapper algorithm for feature selection in classification', *Analytica Chimica Acta*, 1080, pp. 43–54.

Zhang, X., Zhang, Q., Chen, M., Sun, Y., Qin, X. and Li, H. (2018) 'A two-stage feature selection and intelligent fault diagnosis method for rotating machinery using hybrid filter and wrapper method', *Neurocomputing*, 275, pp. 2426-2439.

Zhong, W., Lu, X. and Wu, J. (2017) 'Feature selection for cancer classification using microarray gene expression data', *Biostatistics and Biometrics*, 1(2), pp. 1–7.

Zhou, Y., Jin, R. and Hoi, S.C.H. (2010) 'Exclusive LASSO for multi-task feature selection', *Journal of Machine Learning Research*, 9, pp. 988–995.

LIST OF PUBLICATIONS

Indexed Journal

1. **Hamid, T. M. T. A**., Sallehuddin, R., Yunos, Z. M., & Ali, A. (2019). Ensemble based multi filters algorithm for tumour classification in high dimensional microarray dataset. *International Journal of Advanced Trends in Computer Science and Engineering*, 8(1.6), 116-123. https://doi.org/10.30534/ijatcse/2019/1881.62019. **(Indexed by SCOPUS)**

Indexed Conference Proceedings

1. **Hamid, T. M. T. A**., Sallehuddin, R., & Yunos, Z. M. (2019). Utilization of filter feature selection with Support Vector Machine for tumours classification. In *IOP Conference Series: Materials Science and Engineering*, 551(1), 012062. https://doi.org/10.1088/1757-899X/551/1/012062. **(Indexed by SCOPUS)**