A MODIFIED WHALE OPTIMISATION ALGORITHM FOR EFFICIENT FILTER-BASED FEATURE SELECTION IN HIGH-DIMENSIONAL DATASETS

YAB LI YU

A thesis submitted in
fulfilment of the requirement for the award of the
Degree of Master of Information Technology

Faculty of Computer Science and Information Technology
Universiti Tun Hussein Onn Malaysia

DECEMBER 2022

This thesis is wholeheartedly dedicated to my beloved parents
and my dearest three elder brothers.

ACKNOWLEDGEMENT

The completion of this study could not have been possible without the guidance and support of so many people. Their contributions are truly appreciated and gratefully acknowledged. I would like to express my deepest appreciation and indebtedness to the following:

First and foremost, I would like to thank the Ministry of Education for the research grant from the Fundamental Research Grant Scheme, which supported this study. The research process and research publications would not have been possible without the financial support.

I would like to express my deep gratitude to my supervisor and co-supervisor, Assoc. Prof. Dr. Noorhaniza Wahid and Dr. Rahayu A. Hamid, respectively, for their selfless guidance and willingness to give their time so generously, not only in assisting my academic research professionally but also in personally lifting my spirit from time to time.

My very profound gratitude to my beloved and respected parents, Yab Vain Sin and Lok Kiun, and my elder brothers, Yab Mun Pin, Yab Mun Chon, and Yab Mun Yew, for providing me with unfailing support and continuous encouragement throughout my years of study. Without their wise counsel and sympathetic ears, it would be impossible for me to complete my study.

Moreover, I would like to thank the best senior roommate, Ng Li Mun, for providing me with constant care and encouragement. Not to mention the willingness to share her research and life experiences with me. Also, my heartfelt gratitude to the rest of my housemates and best friends, who provided stimulating discussions, as well as happy distractions, to rest my mind off my research.

ABSTRACT

In a high-dimensional dataset (HDD), reducing the "curse of dimensionality" via feature selection is crucial. Recently, the integration of metaheuristic algorithms in solving feature selection problems had yielded outstanding results, according to literature findings. Hence, this study focused on integrating the Whale Optimisation Algorithm (WOA) for filter-based feature selection in HDDs. However, the WOA is known to have a slow convergence speed issue caused by the control parameter, *a*, which influences the balancing of the exploration and exploitation phases, eventually affecting the searching strategy. Therefore, this research proposed a modified WOA (mWOA) by inversing the control parameter values to allow the mWOA more search spaces during the initial searching phase, which eventually would increase the convergence speed. The proposed mWOA was implemented as the filter-based feature selection in four benchmark medical HDDs, namely Colon, CNS, SMK_CAN_187, and GLI_85. The performance of the proposed mWOA was compared against those of two filter-based feature selection algorithms, namely the original WOA and the Grey Wolf Optimiser (GWO). It was proven that the proposed mWOA outperformed the WOA and the GWO in 3 out of 4 cases (75%) in both best and average execution times when selecting the most relevant features of the HDDs. In addition, the mWOA also outperformed the WOA and the GWO in 8 out of 12 test cases (67%) in classification accuracy when using Decision Tree, Support Vector Machine, and Naïve Bayes classifiers.

ABSTRAK

Pengurangan *curse of dimensionality* dengan kaedah pemilihan ciri adalah suatu perkara yang penting dalam set data berdimensi tinggi (HDD). Kini, penggunaan algoritma metaheuristik dalam menyelesaikan masalah pemilihan ciri telah dapat menghasilkan keputusan yang cemerlang menurut kajian lepas. Oleh itu, kajian ini menumpukan pada algoritma metaheuristik yang popular, iaitu *Whale Optimisation Algorithm* (WOA), untuk pemilihan ciri berasaskan *filter* dalam HDD. Walau bagaimanapun, WOA didapati mempunyai isu kelajuan penumpuan perlahan disebabkan oleh parameter kawalan, *a*, yang mempengaruhi keseimbangan fasa penerokaan dan eksploitasi, yang akhirnya mempengaruhi strategi pencarian ciri. Justeru, penyelidikan ini mencadangkan *modified* WOA (mWOA) dengan menyongsangkan nilai parameter kawalan bagi membolehkan mWOA mempunyai lebih banyak ruang carian dalam fasa pencarian awal yang akhirnya meningkatkan kelajuan penumpuan. Dalam kajian ini, mWOA telah digunakan dalam pemilihan ciri berasaskan kaedah *filter* pada empat set data HDD perubatan, iaitu, Colon, CNS, SMK_CAN_187, dan GLI_85. mWOA telah dibandingkan dengan dua algoritma pemilihan ciri berasaskan kaedah *filter*, iaitu WOA asal dan *Grey Wolf Optimiser* (GWO). Berdasarkan keputusan eksperimen, telah terbukti bahawa prestasi mWOA telah mengatasi WOA asal and GWO di dalam 3 daripada 4 kes (75%) untuk kedua-dua kes terbaik dan kes purata masa pelaksanaan apabila memilih ciri yang paling penting daripada HDD. Selain itu, ketepatan klasifikasi mWOA juga mengatasi WOA dan GWO di dalam 8 daripada 12 kes (67%) apabila menggunakan pengelas-pengelas *Decision Tree*, *Support Vector Machine* dan *Naïve Bayes*.

CONTENT

LIST OF TABLES

LIST OF FIGURES

LIST OF SYMBOLS AND ABBREVIATIONS

DT	-	Decision Tree
GWO	-	Grey Wolf Optimiser
HDD	-	High-Dimensional Data
ML	-	Machine Learning
mWOA	-	Modified Whale Optimisation Algorithm
NB	-	Naïve Bayes
SVM	-	Support Vector Machine
WEKA		Waikato Environment for Knowledge Analysis
WOA	-	Whale Optimisation Algorithm

LIST OF PUBLICATIONS

Yab, L.Y., Wahid, N., Hamid, R.A. (2022). A Modified Whale Optimization Algorithm as Filter-Based Feature Selection for High Dimensional Datasets. *SCDM 2022: Recent Advances in Soft Computing and Data Mining*. Springer International Publishing. pp. 90-100. doi: 10.1007/978-3-031-00828-3_9.

Yab, L.Y., Wahid, N., Hamid, R.A. A Meta-Analysis Survey on the Usage of Meta-Heuristic Algorithms for Feature Selection on High-dimensional Datasets. *IEEE Access*. 2022. vol. 10, pp. 122832-122856. doi: 10.1109/ACCESS.2022.3221194.

CHAPTER 1

INTRODUCTION

1.1 Background of study

For the past decades, data mining has always been the research hotspot for many researchers. Data mining is a broad field of data science, which tries to find patterns and characteristics in a massive quantity of data. It includes regression, clustering, and data classification [1]. Data classification is a fascinating task in data mining, which entails assigning the class label of instances based on a previously trained model [2]. Undeniably, with the rapid development of science and technology, the expansion of datasets is not a new phenomenon anymore. Datasets are getting larger and higher in dimensionality over the years. To further discuss the meaning of high-dimensional datasets (HDDs), one must know the form in which a dataset is usually represented. Datasets are typically interpreted as a matrix, with the row representing instances, while the column representing features. Datasets with a great number of features are categorised as HDDs [3]. A high dimensionality results in unmanageable memory constraints and high training and computing costs, which cause the "curse of dimensionality" [3], [4]. Therefore, there is a need to perform dimensionality reduction to reduce the number of features without compromising the retrieval of useful information in HDDs to ensure good classification performance.

Not all features in HDDs are relevant to provide sufficient information for data classification. These irrelevant features could result in low classification performance. Thus, improving the performance of the classification of HDDs relies on feature

selection to perform dimensionality reduction. Feature selection is a process of selecting the most meaningful features [5]. Feature selection can also be defined as omitting irrelevant and non-essential features in HDDs to not only reduce execution time but also increase the predictive precision of a classifier [6]–[8]. Feature selection has two key competing goals: (1) optimising classification efficiency and (2) minimising the number of features to solve the curse of dimensionality [9]. To balance the trade-off between these two opposing priorities, feature selection can be seen as a multi-objective challenge. Therefore, the pre-processing of data is very important to generate compact yet quality datasets for classification. To put it another way, feature selection aims to choose suitable features that contribute the most to the classification model in order to achieve higher accuracy.

Feature selection can be further categorised into three methods, namely wrapper-based, embedded-based, and filter-based [10]. Wrapper-based feature selection makes use of the strength of the base classifiers to find the best features in a dataset, whereas embedded-based feature selection takes place during model training in the machine learning algorithm [11]. Both wrapper-based and embedded-based methods result in higher execution time due to the intervention of the classifiers in the feature selection process. On the other hand, filter-based feature selection methods rely on mutual information in the dataset and rank its features by generating a score for each feature, independent of the classification model [11]. It is worth mentioning that wrapper-based methods are computationally less feasible for HDDs due to the higher execution time by the classification model [4]. As for embedded-based methods, these require certain predictive models, whereas filter-based methods can be combined with any kind of predictive model and are fast when calculating the HDDs [4]. Among these three methods, it is noticeable that filter-based feature selection selects a subset of features without using any learning algorithm, and thus it is relatively faster than wrapper-based methods and useful in HDDs. Not only that, filter-based methods have low complexity among all types of feature selection and are compatible with diverse datasets, including HDDs [4], [11].

Metaheuristic optimisation algorithms proposed by researchers have been used to simplify classification and solve feature selection issues for decades. In metaheuristic algorithms, exploitation and exploration are the two fundamental components that control the searching mechanism to obtain the optimal solution [12].

In exploration, the optimiser must contain operators to explore the search space globally, and the motions are randomised as much as possible during this phase. On the other hand, exploitation is the process of investigating thoroughly the promising section of the search space found during exploration [13]. Some examples of metaheuristic optimisation algorithms are Gravitational Search Algorithm (GSA) [14], Ant Colony Optimisation (ACO) algorithm [15], Grey Wolf Optimiser (GWO) [16], Ant Lion Optimiser (ALO) [17], Particle Swarm Optimisation (PSO) [18], and Whale Optimisation Algorithm (WOA) [13]. Many researchers have employed these metaheuristic algorithms in various domains, such as solving power system problems in electrical engineering [19], applying multivariate data clustering [20], solving electromagnetic problems [21], performing data classification [22], and solving feature selection problems [11], [23], [24]. Among these algorithms, the WOA shows its strength in balancing exploration and exploitation, making it the top optimiser as compared with the other aforementioned metaheuristic algorithms [12], [23], [24].

The WOA is a swarm-based nature-inspired metaheuristic algorithm that mimics the biological behaviour of humpback whales to solve optimisation problems [13]. The algorithm consists of three parts, which are encircling prey, spiral updating of position, and searching for prey. The first two parts are implemented in the exploitation phase, while the latter part is done randomly in the exploration phase. The WOA is widely used in various areas, such as processing diabetes data [25], predicting traffic congestion rates [26], and optimising feature selection in medical datasets [23]. The WOA has proven itself to outperform the aforementioned algorithms in feature selection with a better ability to search for optimal features, leading to maximum classification accuracy [12].

It is worth noticing that researchers using wrapper-based feature selection with the WOA showed high execution times [23], [24], [27]. The WOA has also been applied in filter-based methods in various research works and was able to produce the best accuracy when 50% of features were omitted [11]. However, there are some drawbacks of the WOA yet to be solved. For instance, the performance of the WOA is affected by the convergence speed [12], [28]–[31], influencing the feature selection process by being unable to select the most relevant features or having too long of an execution time. Therefore, this research aimed to overcome the convergence speed

issue of the original WOA and improve the algorithm's capability of selecting the most relevant features in a shorter execution time.

1.2 Problem statement

The convergence speed plays an important role in the performance of metaheuristic algorithms. It indicates how fast an algorithm converges to the optimum solution. If the algorithm converges to the optimum too quickly, it is likely that the best solution might be overlooked, which reduces classification accuracy [32]. Likewise, when the algorithm converges to the optimum too slowly, then the algorithm is not performing very well because it takes too long to find the best solutions. This needs to be avoided, especially when the data are huge.

However, slow convergence speed issue is found in the literature for filter-based feature selection methods with metaheuristic algorithms. For instance, it is worth mentioning that similar to other metaheuristic algorithms, WOA still face the slow convergence speed issue [28]. Based on the literature, a filter-based feature selection method using WOA with Mutual Congestion (WOA-MC) [11] has a slow convergence speed issue in HDD that is yet to be solved.

In the WOA, the convergence speed depends on a single control parameter, *a*, which has a large effect on the WOA's performance, such as balancing between exploration and exploitation [12], [29]. This is because *a* is used to generate the value of the coefficient vector, $\vec{A}$, which then affects the equations of position updating in the phases of encircling prey, bubble-net attacking (exploitation), and searching for prey (exploration). The control parameter, *a*, affects the algorithm's convergence in both local and global searching strategies. Its function determines the distance between search agents and the likelihood of position changing to look for solutions in the search space, which eventually results in the convergence speed of the algorithm [33]. As a result, the WOA has a slow rate of convergence throughout both the exploration and exploitation phases [30], [31]. Thus, the process of controlling the parameter needs to be improved in order to achieve a balance between these phases [12], [34].

As inspired by a Binary Grey Wolf Optimizer (BGWO) for wrapper-based feature selection [33], whereby the control parameter in BGWO was altered to linearly

increase to improve the slow convergence issue; it is suggested that the control parameter might be able to determine the convergence speed in filter-based WOA too. Hence, the issue of slow convergence speed in WOA is chosen in this study, by changing the control parameter to produce linearly increasing values over iteration.

Therefore, in this study, a modified WOA (mWOA) was proposed to improve the convergence speed for better performance of the searching mechanism for filter-based feature selection in HDDs. This proposed method was expected to perform better than the original WOA in balancing exploration and exploitation, as well as in improving the classification of HDDs.

1.3 Objectives of study

The aims of the research were twofold: to design an algorithm that selects the most relevant features in HDDs with a shorter execution time and to obtain a better classification performance by solving the convergence speed issue in the WOA. To fulfil the aims of the research, the following objectives were set:

i. To propose a modified WOA (mWOA) by inversing the values of the control parameter, *a*, to tackle the slow convergence speed issue during exploration and exploitation.

ii. To implement the mWOA as a filter-based feature selection method to select the most relevant features in benchmark medical HDDs.

iii. To evaluate the performance of the mWOA against those of two other filter-based feature selection methods, namely the original WOA and the GWO.

1.4 Scope of study

This research focused on filter-based feature selection using metaheuristic algorithms. The improvement made by the proposed mWOA was limited to solving the convergence speed issue of the original WOA.

In this study, medical HDDs were selected due to the increasing use of data mining in the area of medical diagnosis [35]–[39]. Four well-known benchmark medical HDDs of the binary class were used to validate the proposed mWOA in this study, namely Colon, Central Nervous System (CNS), GLI_85, and SMK_CAN_187 [11], [40]. Binary-class HDDs were employed because the fitness function for filter-based feature selection using the WOA required calculating the Euclidean distance between two classes, as adopted from [11].

To perform feature selection and classification, these four HDDs were tested using the WOA, the GWO, and the proposed mWOA. As suggested by Nematzadeh *et al*. [11], the feature selection discard rate was set to 50%; hence, only half of the features were selected by each algorithm. The performance of the proposed mWOA was compared against those of the WOA and the GWO using two evaluation criteria to prove the effectiveness of the modified control parameter.

The first evaluation criteria was the execution time taken to select relevant features. Specifically, each algorithm's best, average, and worst execution times were evaluated. Besides execution time, the performances of the algorithms in classifying the selected features were also evaluated. There were three classifiers used: Decision Tree, Naïve Bayes, and Support Vector Machine. These classifiers are commonly used for classifying medical HDDs [11], [40]. The classification performances of these classifiers were evaluated in terms of accuracy, specificity, and sensitivity.

1.5 Significance of study

Metaheuristic algorithms have contributed to various areas, such as solving problems in engineering, optimisation, and feature selection. Hence, using an efficient metaheuristic algorithm could significantly improve an application's performance. The WOA, despite being one of the well-performing optimisation algorithms, has a limitation in obtaining a faster convergence speed, an issue that has been studied by previous researchers. Therefore, this study proposed a modified WOA (mWOA), an algorithm that is well balanced between exploration and exploitation phases and thus is able to achieve a faster convergence speed, by modifying the control parameter, *a*.

The proposed mWOA could be useful for researchers who are interested in integrating an optimisation algorithm in the mentioned areas.

Data mining of HDDs relies on feature selection for dimensionality reduction to avoid the curse of dimensionality. Hence, an efficient approach for feature selection using metaheuristic algorithms could significantly improve the data mining of HDDs. In this study, the proposed mWOA was implemented as the filter-based feature selection in binary-class medical HDDs. The proposed mWOA was expected to select the most relevant features in a shorter execution time and obtain a better classification performance. This filter-based feature selection method could be useful for the feature selection of especially, but not limited to, medical HDDs.

As reported in [41], in the year 2021, there are 48,639 new cancer cases recorded in Malaysia, and this number is expected to get doubled by the year 2040. Unfortunately, the oncology field in Malaysia is still relatively new as compared to other medical disciplines, and expert oncologists are only available in big general hospitals or private facilities in major cities [41]. Therefore, it would be useful if the proposed mWOA could perform feature selection on cancerous datasets which might increase the chances of getting a more accurate diagnosis. With the fast pace of data mining advancement in medical diagnosis, it is inevitable that the medical datasets are becoming bigger and bigger to hold more useful information. Therefore, tackling the curse of dimensionality in medical HDD with the proposed feature selection method is a prioritised task. This could contribute to researchers in the oncology discipline to obtain accurate medical diagnoses more efficiently. It is also hoped that the improved mWOA would be beneficial to select meaningful features with shorter execution time yet able to contribute higher accuracy while classifying tumor and normal genes in these datasets.

1.6 Thesis outline

This thesis dissertation is organised into five chapters, and each chapter covers several subsections. Chapter 1 introduced the background of the study and the problem statement that motivated the research work, as well as the objectives, scope, and significance of the study.

Chapter 2 presents the literature review covering high-dimensional data, feature selection, metaheuristic algorithms for feature selection, and a systematic review on feature selection using metaheuristic approaches in HDDs, as well as discussions on the Whale Optimisation Algorithm, the WOA for feature selection, the Grey Wolf Optimiser, and the research gap.

In Chapter 3, the overall methodology to conduct the research, which involved data acquisition, mWOA formulation, feature selection, data classification, and evaluation metrics, was demonstrated. The discussion on the experiment setup for feature selection in MATLAB and for classification in WEKA is also covered in this chapter.

In Chapter 4, the experiment results and findings are discussed. The chapter begins with comparisons of execution times for feature selection. The proposed mWOA's best, average, and worst execution times were compared against those of the WOA and the GWO. Besides that, classification performances in terms of accuracy, sensitivity, and specificity are discussed in this chapter.

Chapter 5 concludes the thesis, and the research's novelty and contribution are explained. Suggestions and recommendations for future works are also provided in this chapter.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

In this chapter, related research works are discussed to give a deeper understanding of the background of this study. The topics of high-dimensional datasets (HDDs), feature selection, metaheuristic approaches, a systematic review of feature selection that used metaheuristic approaches in HDDs, the Whale Optimisation Algorithm (WOA), and the WOA's applications are all included. The research gap and the justification for each decision are presented with solid evidence at the end of this chapter.

2.2 High-dimensional datasets

Over the past few years, quintillion bytes of data are created every day [42]. As a result, vast volumes of data with very high dimensions have arisen in various machine learning (ML) applications, including data mining [43]. Data mining often deals with a wide variety of distinct datasets. Due to the increasing number of data and complexity of datasets, extracting usable information from massive quantities of irrelevant information in datasets has become more important.

Typically, datasets are in the form of matrices, where the row represents instances, while the column represents features [3]. HDDs are datasets that have a large number of features and are thus more complex. Low-dimensional datasets, in contrast,

have fewer features and are narrower in size. Figure 2.1 illustrates the visualisation of data dimension.

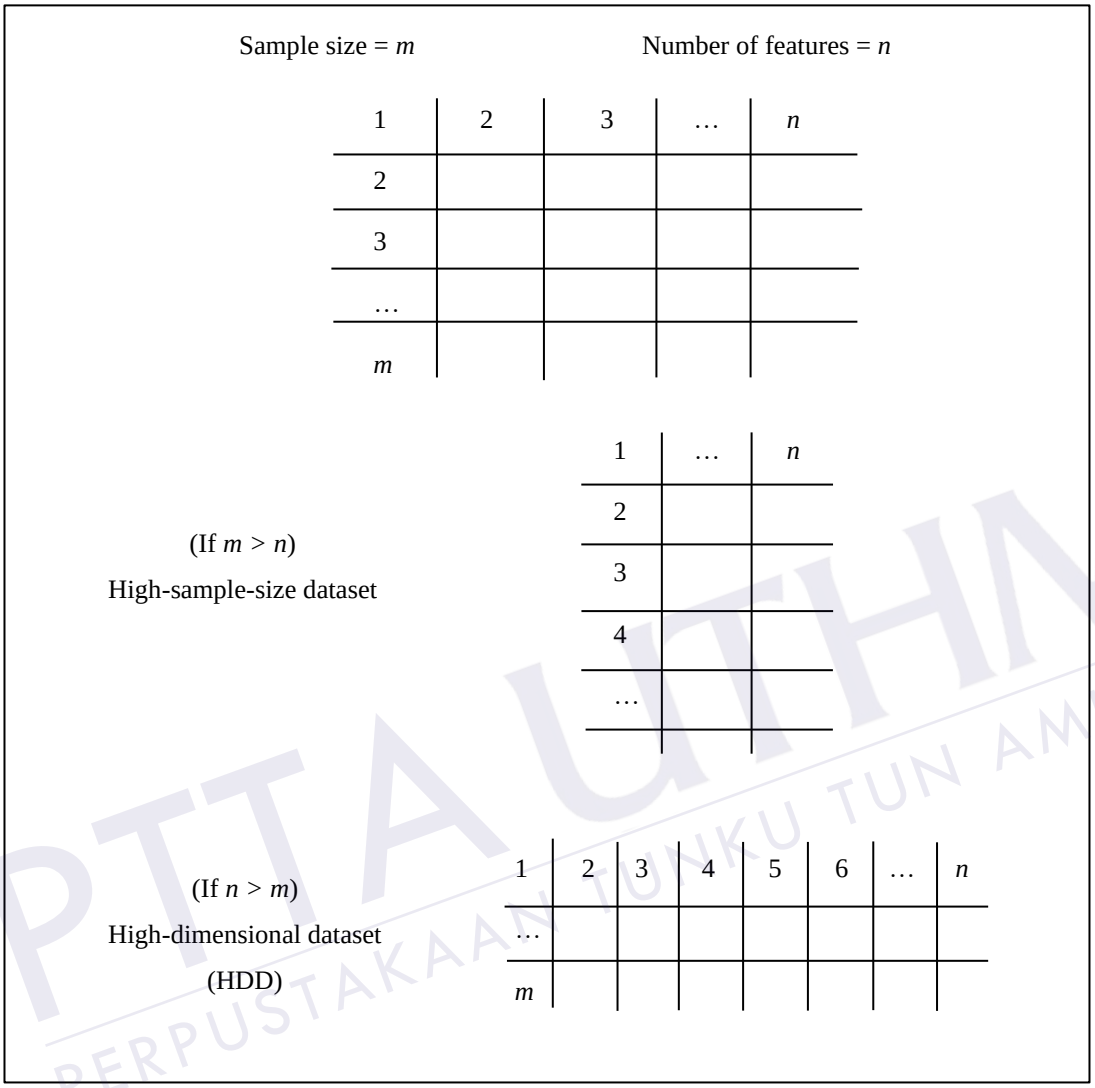

Figure 2.1: Visualisation of data dimension

The number of features and the sample size of a dataset are considered in order to categorise it as a high-sample-size dataset, or HDD. Letting *m* be the size of the sample and *n* as the number of features, a dataset has a high sample size if $m > n$. In other words, a dataset is considered as a high-sample-size dataset if its sample size is greater than the number of features. On the contrary, if its sample size is smaller than the number of features, the dataset is considered as an HDD. Nowadays, HDDs are becoming more prevalent in various areas, including text recognition, medical imaging, genetic microarrays, finance, face recognition, and chemometrics [3].

REFERENCES

[1] X.-S. Yang, *Introduction to Algorithms for Data Mining and Machine Learning*. Elsevier, 2019.

[2] R. Alazaidah, M. A. Almaiah, and M. Al-Luwaici, "Associative Classification In Multi-label Classification: An Investigative Study," *Jordanian J. Comput. Inf. Technol.*, vol. 7, no. 2, pp. 166–179, 2021.

[3] V. Bolón-Canedo, N. Sánchez-Maroño, and A. Alonso-Betanzos, *Feature Selection for High-Dimensional Data*. Springer International Publishing, 2015.

[4] A. Bommert, X. Sun, B. Bischl, J. Rahnenführer, and M. Lang, "Benchmark for filter methods for feature selection in high-dimensional classification data," *Comput. Stat. Data Anal.*, vol. 143, p. 106839, 2020.

[5] O. Duncan and T. Sherer, "Feature Selection (Data Mining)," *Microsoft*, 2018. [Online]. Available: https://docs.microsoft.com/en-us/analysis-services/data-mining/feature-selection-data-mining?view=asallproducts-allversions. [Accessed: 01-May-2021].

[6] B. Zhang and P. Cao, "Classification of high dimensional biomedical data based on feature selection using redundant removal," *PLoS One*, vol. 14, no. 4, pp. 1–19, 2019.

[7] A. Veeraswamy and A. M. Babu, "Classification of High Dimensional Data Using Filtration Attribute Evaluation Feature Selection Method of Data mining," *4th Int. Conf. Electr. Electron. Commun. Comput. Technol. Optim. Tech. ICEECCOT 2019*, pp. 8–12, 2019.

[8] K. S. Adewole *et al.*, "Hybrid Feature Selection Framework For Sentiment Analysis On Large Corpora," *Jordanian J. Comput. Inf. Technol.*, vol. 07, no. 02, pp. 15–33, 2021.

[9] Q. Al-Tashi, S. J. Abdulkadir, H. M. Rais, S. Mirjalili, and H. Alhussian, "Approaches to Multi-Objective Feature Selection: A Systematic Literature Review," *IEEE Access*, vol. 8, pp. 125076–125096, 2020.

[10] L. Hu, W. Gao, K. Zhao, P. Zhang, and F. Wang, "Feature selection considering two types of feature relevancy and feature interdependency," *Expert Syst. Appl.*, vol. 93, pp. 423–434, 2018.

[11] H. Nematzadeh, R. Enayatifar, M. Mahmud, and E. Akbari, "Frequency based feature selection method using whale algorithm," *Genomics*, vol. 111, no. 6, pp. 1946–1955, 2019.

[12] H. M. Mohammed, S. U. Umar, and T. A. Rashid, "A systematic and meta-analysis survey of whale optimization algorithm," *Comput. Intell. Neurosci.*, vol. 2019, 2019.

[13] S. Mirjalili and A. Lewis, "The Whale Optimization Algorithm," *Adv. Eng. Softw.*, vol. 95, pp. 51–67, 2016.

[14] E. Rashedi, H. Nezamabadi-pour, and S. Saryazdi, "GSA: A Gravitational Search Algorithm," *Inf. Sci.*, vol. 179, no. 13, pp. 2232–2248, 2009.

[15] M. Dorigo, M. Birattari, and T. Stützle, "Ant Colony Optimization," *IEEE Comput. Intell. Mag.*, vol. 1, no. 4, pp. 28–39, 2006.

[16] S. Mirjalili, S. M. Mirjalili, and A. Lewis, "Grey Wolf Optimizer," *Adv. Eng. Softw.*, vol. 69, pp. 46–61, 2014.

[17] S. Mirjalili, "The Ant Lion Optimizer," *Adv. Eng. Softw.*, vol. 83, pp. 80–98, 2015.

[18] J. Kennedy and R. Eberhart, "Particle Swarm Optimization," *Proc. ICNN'95 - Int. Conf. Neural Networks*, pp. 1942–1948, 1995.

[19] G. Chen, L. Liu, and S. Huang, "Enhanced GSA-based optimization for minimization of power losses in power system," *Math. Probl. Eng.*, vol. 2015, 2015.

[20] W. Gao, "Improved ant colony clustering algorithm and its performance study," *Comput. Intell. Neurosci.*, vol. 2016, 2016.

[21] X. Li and K. M. Luk, "The Grey Wolf Optimizer and Its Applications in Electromagnetics," *IEEE Trans. Antennas Propag.*, vol. 68, no. 3, pp. 2186–2197, 2020.

[22] A. S. Assiri, A. G. Hussien, and M. Amin, "Ant lion optimization: Variants, hybrids, and applications," *IEEE Access*, vol. 8, no. April, pp. 77746–77764, 2020.

[23] M. Mafarja and S. Mirjalili, "Whale optimization approaches for wrapper

feature selection," *Appl. Soft Comput.*, vol. 62, pp. 441–453, 2018.

[24] M. M. Mafarja and S. Mirjalili, "Hybrid Whale Optimization Algorithm with simulated annealing for feature selection," *Neurocomputing*, vol. 260, pp. 302–312, 2017.

[25] Rajeshkumar and Kousalya, "Diabetes Data Classification Using Whale Optimization Algorithm and Backpropagation Neural Network," *Int. Res. J. Pharm.*, vol. 8, no. 11, pp. 219–222, 2017.

[26] B. Sony, A. Chakravarti, and M. M. Reddy, "Traffic congestion detection using whale optimization algorithm and multi-support vector machine," *Int. J. Recent Technol. Eng.*, vol. 7, no. 6C2, pp. 589–593, 2019.

[27] M. Sharawi, H. M. Zawbaa, and E. Emary, "Feature selection approach based on whale optimization algorithm," *9th Int. Conf. Adv. Comput. Intell. ICACI 2017*, pp. 163–168, 2017.

[28] P. Niu, S. Niu, N. liu, and L. Chang, "The defect of the Grey Wolf optimization algorithm and its verification method," *Knowledge-Based Syst.*, vol. 171, pp. 37–43, 2019.

[29] M. Zhong and W. Long, "Whale optimization algorithm with nonlinear control parameter," *MATEC Web Conf.*, vol. 139, pp. 1–5, 2017.

[30] R. K. Saidala and N. Devarakonda, "Improved whale optimization algorithm case study: Clinical data of anaemic pregnant woman," *Adv. Intell. Syst. Comput.*, vol. 542, pp. 271–281, 2018.

[31] L. M. Pecora and T. L. Carroll, "Synchronization of chaotic systems," *Chaos*, vol. 25, no. 9, 2015.

[32] S. Qian, Y. Shi, H. Wu, and S. Shang, "An Improved Hybrid Feature Selection Algorithm for Electric Charge Recovery Risk," *Math. Probl. Eng.*, vol. 2020, 2020.

[33] P. Hu, J. S. Pan, and S. C. Chu, "Improved Binary Grey Wolf Optimizer and Its application for feature selection," *Knowledge-Based Syst.*, vol. 195, p. 105746, 2020.

[34] M. Abdel-Basset, G. Manogaran, D. El-Shahat, and S. Mirjalili, "A hybrid whale optimization algorithm based on local search strategy for the permutation flow shop scheduling problem," *Futur. Gener. Comput. Syst.*, vol. 85, no. March, pp. 129–145, 2021.

[35] P. Jaganathan and R. Kuppuchamy, “A threshold fuzzy entropy based feature selection for medical database classification,” *Comput. Biol. Med.*, vol. 43, no. 12, pp. 2222–2229, 2013.

[36] S. Shilaskar and A. Ghatol, “Feature selection for medical diagnosis: Evaluation for cardiovascular diseases,” *Expert Syst. Appl.*, vol. 40, no. 10, pp. 4146–4153, 2013.

[37] S. Nagpal, S. Arora, S. Dey, and S. Shreya, “Feature Selection using Gravitational Search Algorithm for Biomedical Data,” *Procedia Comput. Sci.*, vol. 115, pp. 258–265, 2017.

[38] Q. Liu, Q. Gu, and Z. Wu, “Feature selection method based on support vector machine and shape analysis for high-throughput medical data,” *Comput. Biol. Med.*, vol. 91, no. March, pp. 103–111, 2017.

[39] Y. Peng, Z. Wu, and J. Jiang, “A novel feature selection approach for biomedical data classification,” *J. Biomed. Inform.*, vol. 43, no. 1, pp. 15–23, 2010.

[40] Z. Sadeghian, E. Akbari, and H. Nematzadeh, “A hybrid feature selection method based on information theory and binary butterfly optimization algorithm,” *Eng. Appl. Artif. Intell.*, vol. 97, no. February 2020, p. 104079, 2021.

[41] M. M. Yusof and W. Z. W. Ishak, “Cancer in My Community: Addressing Increasing Cancer Cases in Malaysia,” *ASCO Cancer.net*, 2022. [Online]. Available: https://www.cancer.net/blog/2022-02/cancer-my-community-addressing-increasing-cancer-cases-malaysia. [Accessed: 09-Nov-2022].

[42] Y. Zhai, Y. S. Ong, and I. W. Tsang, “The emerging ‘Big dimensionality,’” *IEEE Comput. Intell. Mag.*, vol. 9, no. 3, pp. 14–26, 2014.

[43] M. Tan, I. W. Tsang, and L. Wang, “Towards ultrahigh dimensional feature selection for big data,” *J. Mach. Learn. Res.*, vol. 15, pp. 1371–1429, 2014.

[44] R. C. Thom de Souza, C. A. de Macedo, L. dos Santos Coelho, J. Pierezan, and V. C. Mariani, “Binary coyote optimization algorithm for feature selection,” *Pattern Recognit.*, vol. 107, p. 107470, 2020.

[45] Y. Zhang, X. F. Song, and D. W. Gong, “A return-cost-based binary firefly algorithm for feature selection,” *Inf. Sci.*, vol. 418–419, pp. 561–574, 2017.

[46] J. H. Holland, “Genetic algorithms,” *Sci. Am.*, vol. 267, no. 1, pp. 66–72, 1992.

[47] I. Rechenberg, “Evolutionsstrategien BT - Simulationsmethoden in der Medizin

und Biologie," B. Schneider and U. Ranft, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1978, pp. 83–114.

[48] J. R. Koza and R. Poli, "Genetic programming," in *Search Methodologies: Introductory Tutorials in Optimization and Decision Support Techniques*, Springer US, 2005, pp. 127–164.

[49] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, "Optimization by simulated annealing," *Science (80-.).*, vol. 220, no. 4598, pp. 671–680, 1983.

[50] A. Hatamlou, "Black hole: A new heuristic optimization approach for data clustering," *Inf. Sci.*, vol. 222, pp. 175–184, Feb. 2013.

[51] R. V. Rao, V. J. Savsani, and D. P. Vakharia, "Teaching-Learning-Based Optimization: An optimization method for continuous non-linear large scale problems," *Inf. Sci.*, vol. 183, no. 1, pp. 1–15, Jan. 2012.

[52] A. H. Kashan, "League Championship Algorithm: A new algorithm for numerical function optimization," in *SoCPaR 2009 - Soft Computing and Pattern Recognition*, 2009, pp. 43–48.

[53] C. Dai, W. Chen, Y. Song, and Y. Zhu, "Seeker optimization algorithm: A novel stochastic search algorithm for global numerical optimization," *J. Syst. Eng. Electron.*, vol. 21, no. 2, pp. 300–311, Apr. 2010.

[54] F. Moslehi and A. Haeri, "A novel hybrid wrapper–filter approach based on genetic algorithm, particle swarm optimization for feature subset selection," *J. Ambient Intell. Humaniz. Comput.*, vol. 11, no. 3, pp. 1105–1127, 2020.

[55] M. Taradeh *et al.*, "An evolutionary gravitational search-based feature selection," *Inf. Sci.*, vol. 497, pp. 219–239, 2019.

[56] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary ant lion approaches for feature selection," *Neurocomputing*, vol. 213, pp. 54–65, 2016.

[57] S. Arora and P. Anand, "Binary butterfly optimization approaches for feature selection," *Expert Syst. Appl.*, vol. 116, pp. 147–160, 2019.

[58] M. Mafarja *et al.*, "Binary dragonfly optimization for feature selection using time-varying transfer functions," *Knowledge-Based Syst.*, vol. 161, no. August, pp. 185–204, 2018.

[59] M. Mafarja, I. Aljarah, H. Faris, A. I. Hammouri, A. M. Al-Zoubi, and S. Mirjalili, "Binary grasshopper optimisation algorithm approaches for feature selection problems," *Expert Syst. Appl.*, vol. 117, pp. 267–286, 2019.

[60] E. Emary, H. M. Zawbaa, and A. E. Hassanien, "Binary grey wolf optimization approaches for feature selection," *Neurocomputing*, vol. 172, pp. 371–381, 2016.

[61] H. Hichem, M. Elkamel, M. Rafik, M. T. Mesaaoud, and C. Ouahiba, "A new binary grasshopper optimization algorithm for feature selection problem," *J. King Saud Univ. - Comput. Inf. Sci.*, vol. 34, no. 2, pp. 316–328, 2019.

[62] M. Alweshah, S. Al Khalaileh, B. B. Gupta, A. Almomani, A. I. Hammouri, and M. A. Al-Betar, "The monarch butterfly optimization algorithm for solving feature selection problems," *Neural Comput. Appl.*, vol. 0, 2020.

[63] J. Nasiri, F. M. Khiyabani, and A. Yoshise, "A whale optimization algorithm (WOA) approach for clustering," *Cogent Math. Stat.*, vol. 5, no. 1, p. 1483565, 2018.

[64] M. A. Ahmed, I. Ahsan, and M. Abbas, "Systematic literature review: Ingenious software project management while narrowing the impact aspect," *Proc. 2016 Res. Adapt. Converg. Syst. RACS 2016*, pp. 165–168, 2016.

[65] D. H. Wolpert and W. G. Macready, "No free lunch theorems for optimization," *IEEE Trans. Evol. Comput.*, vol. 1, no. 1, pp. 67–82, 1997.

[66] F. S. Gharehchopogh and H. Gholizadeh, "A comprehensive survey: Whale Optimization Algorithm and its applications," *Swarm Evol. Comput.*, vol. 48, no. November 2018, pp. 1–24, 2019.

[67] V. Ho-Huu, T. Nguyen-Thoi, M. H. Nguyen-Thoi, and L. Le-Anh, "An improved constrained differential evolution using discrete variables (D-ICDE) for layout optimization of truss structures," *Expert Syst. Appl.*, vol. 42, no. 20, pp. 7057–7069, 2015.

[68] Ö. Çelik, "A Research on Machine Learning Methods and Its Applications," *J. Educ. Technol. Online Learn.*, vol. 1, no. 3, pp. 25–40, 2018.

[69] H. Mohammadzadeh and F. S. Gharehchopogh, "A novel hybrid whale optimization algorithm with flower pollination algorithm for feature selection: Case study Email spam detection," *Comput. Intell.*, vol. 37, no. 1, pp. 176–209, 2021.

[70] M. Ghosh, R. Guha, R. Sarkar, and A. Abraham, "A wrapper-filter feature selection technique based on ant colony optimization," *Neural Comput. Appl.*, vol. 32, no. 12, pp. 7839–7857, 2020.

[71] U. of Waikato, "Weka," *waikato.github.io*, 2014. [Online]. Available: https://waikato.github.io/weka-wiki/. [Accessed: 24-Jun-2021].

[72] Y. Zheng *et al.*, "A novel hybrid algorithm for feature selection," *Pers. Ubiquitous Comput.*, vol. 22, no. 5–6, pp. 971–985, 2018.

[73] Jason Brownlee, "How To Use Classification Machine Learning Algorithms in Weka," *machinelearningmastery.com*, 2016. [Online]. Available: https://machinelearningmastery.com/use-classification-machine-learning-algorithms-weka/. [Accessed: 21-Jul-2022].

[74] T. A. Munandar and Sumiati, "The classification of cropping patterns based on regional climate classification using decision tree approach," *J. Comput. Sci.*, vol. 13, no. 9, pp. 408–415, 2017.

[75] H. Nematzadeh, "Repository of Frequency-based feature selection method using whale algorithm," *github.com*, 2018. [Online]. Available: https://github.com/hnematzadeh/Frequency-based-feature-selection-method-using-whale-algorithm. [Accessed: 17-Jun-2021].

[76] M. Hammami, S. Bechikh, C. C. Hung, and L. Ben Said, "A Multi-objective hybrid filter-wrapper evolutionary approach for feature selection," *Memetic Comput.*, vol. 11, no. 2, pp. 193–208, 2019.

VITA

The author was born on April 11, 1996, in Johor, Malaysia. Her early education began with Sekolah Jenis Kebangsaan (C) Foon Yew 4 in 2003 for her primary education. She continued her secondary school at Sekolah Menengah Kebangsaan Taman Pelangi in 2009 and she studied science stream during her high school. After 5 years of secondary education, the author entered Sekolah Menengah Kebangsaan Dato' Jaafar for Pre-University programmes in 2014 where she further studied science stream. Before going to the university, the author spent a few months working part-time jobs and has broaden her horizons by tutoring, working as general clerk in a furniture warehouse, and digital marketing in a technology company. Consequently, she became more passionate in sharing knowledges and experiences with others and grew interests in multimedia and computing skills. The author then pursued her degree in Bachelor of Computer Science (Multimedia Computing) with Honours at Universiti Tun Hussein Onn Malaysia in 2016. Her final year project received an award for the best project in multimedia computing department in 2019. In the same year, she has taken the industrial training at Ramatex Malaysia where she explored web application and system development. After she received her bachelor's degree in 2020, she continued her employment for another year at the company. In 2021, the author further pursued her study on Master in Information Technology at the same university and she began her research in the area of data mining and soft computing.