IDENTIFICATION OF PATHWAY AND GENE MARKERS USING ENHANCED DIRECTED RANDOM WALK FOR MULTICLASS CANCER EXPRESSION DATA

NIES HUI WEN

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

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

JUNE 2020

DEDICATION

This thesis is dedicated to my parents, brother, supervisors, godparents, friends, and late grandma, who taught me that “learning from mistakes”, “don’t speak unless you can improve on the silence”, “self-control is a key to achieve success”, and “always remember who lent you a helping hand before”.

ACKNOWLEDGEMENT

In preparing this thesis, I received non-stop caring, love, and support from many people. At the same time, I was in contact with researchers and academicians. They have contributed towards my understanding and thoughts. First, I wish to convey my deepest gratitude to my supervisors, Dr. Zalmiyah Zakaria, Dr. Chan Weng Howe, and Professor Dr. Mohd Saberi Mohamad (from Universiti Malaysia Kelantan [UMK], Malaysia) for their encouragement, guidance, and critics. Also, I am thankful to the head of Artificial Intelligence and Bioinformatics Research Group (AIBIG), Dr. Azurah A Samah for her guidance and advice. Additionally, a special thanks to Professor Dr Nazar Zaki (from United Arab Emirate University [UAEU], UAE), Professor Dr Juan Manuel Corchado (from Universidad de Salamanca [USAL], Spain), and Professor Dr Javier Bajo (from Universidad Politécnica de Madrid, Spain) for sharing their ideas and supports. I am also indebted to Universiti Teknologi Malaysia (UTM) Zamalah's scholarship for funding my Ph.D. study. Librarians at UTM and the Google Scholar also deserve special thanks for supplying the relevant literature. My sincere appreciation also extends to those who have helped me on various occasions. Unfortunately, it is impossible to list all of them here. Other than that, I am grateful to all my family members too. Without all of you, I will not reach this stage, and this thesis would not have been presented here.

ABSTRACT

Cancer markers play a significant role in the diagnosis of the origin of cancers and in the detection of cancers from initial treatments. This is a challenging task owing to the heterogeneity nature of cancers. Identification of these markers could help in improving the survival rate of cancer patients, in which dedicated treatment can be provided according to the diagnosis or even prevention. Previous investigations show that the use of pathway topology information could help in the detection of cancer markers from gene expression. Such analysis reduces its complexity from thousands of genes to a few hundreds of pathways. However, most of the existing methods group different cancer subtypes into just disease samples, and consider all pathways contribute equally in the analysis process. Meanwhile, the interaction between multiple genes and the genes with missing edges has been ignored in several other methods, and hence could lead to the poor performance of the identification of cancer markers from gene expression. Thus, this research proposes enhanced directed random walk to identify pathway and gene markers for multiclass cancer gene expression data. Firstly, an improved pathway selection with analysis of variances (ANOVA) that enables the consideration of multiple cancer subtypes is performed, and subsequently the integration of k-mean clustering and average silhouette method in the directed random walk that considers the interaction of multiple genes is also conducted. The proposed methods are tested on benchmark gene expression datasets (breast, lung, and skin cancers) and biological pathways. The performance of the proposed methods is then measured and compared in terms of classification accuracy and area under the receiver operating characteristics curve (AUC). The results indicate that the proposed methods are able to identify a list of pathway and gene markers from the datasets with better classification accuracy and AUC. The proposed methods have improved the classification performance in the range of between 1% and 35% compared with existing methods. Cell cycle and p53 signaling pathway were found significantly associated with breast, lung, and skin cancers, while the cell cycle was highly enriched with squamous cell carcinoma and adenocarcinoma.

ABSTRAK

Penanda kanser memainkan peranan penting dalam mengesan tanda-tanda penyakit kanser dan membolehkan rawatan dilakukan pada peringkat awal. Tugas ini mencabar disebabkan oleh keunikan sifat kanser itu sendiri. Pengenalpastian penanda ini boleh membantu meningkatkan kadar survival pesakit kanser apabila rawatan bersesuaian dapat diberikan dan usaha pencegahan dipertingkatkan. Kajian terdahulu menunjukkan bahawa penggunaan maklumat topologi dan laluan dapat membantu dalam mengesan penanda kanser dari ekspresi gen. Analisis ini dapat mengurangkan kerumitan sumber maklumatnya dari ribuan gen kepada ratusan laluan. Walau bagaimanapun, kebanyakan kaedah sedia ada mengkelaskan semua jenis kanser yang berbeza kepada satu petunjuk penyakit sahaja dan menganggap semua laluan adalah sama. Manakala dalam beberapa kaedah lain, interaksi antara gen dan gen yang terpisah daripada rangkaian telah diabaikan. Ini boleh menyebabkan kemerosotan prestasi pengenalpastian penanda kanser daripada ekspresi gen. Justeru, kajian ini mencadangkan kaedah yang dipertingkatkan bagi perjalanan rawak terarah untuk mengenalpasti gen dan laluan bermaklumat dari data ekspresi gen yang berasaskan pelbagai kelas kanser. Pertama, pemilihan laluan yang bertambah baik dilakukan dengan menggunakan analisis varians yang membolehkan pertimbangan pelbagai kelas kanser. Kedua, pengintegrasian pengelompokan k-means dan kaedah siluet purata dalam perjalanan rawak terarah yang mempertimbangkan interaksi pelbagai gen pula dilakukan. Kaedah yang dicadangkan telah diuji pada kumpulan data penanda aras iaitu ekspresi gen (kanser payudara, paru-paru, dan kulit) dan laluan biologi. Prestasi pengkelasan dari segi ketepatan dan luas di bawah lengkung berasaskan penerima operasi sifat yang dapat dicapai oleh kaedah yang dicadangkan ini telah diukur dan dibandingkan. Dapatan kajian menunjukkan bahawa kaedah yang dicadangkan dapat mengenalpasti senarai penanda laluan dan gen dengan ketepatan pengkelasan dan AUC yang lebih baik. Kaedah yang dicadangkan telah meningkatkan prestasi pengelasan dalam julat antara 1% hingga 35% berbanding dengan kaedah lain. Kitaran sel dan laluan isyarat p53 telah didapati secara ketara berkaitan dengan kanser payudara, paru-paru, dan kulit, sementara kitaran sel diperkayakan dengan karsinoma sel skuamus dan adenokarsinoma.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AC	-	Adenocarcinoma
AFS	-	ANOVA-based Feature Set
AGNES	-	Agglomerative Nesting
AHC	-	Agglomerative Hierarchical Clustering
ANOVA	-	Analysis of Variance
AUC	-	Area Under the Receiver Operating Characteristics Curve
AVA	-	All-Versus-All
CCND1	-	Cyclin D1
CePa	-	Centrality-Based Pathway Enrichment
CLARANS	-	Clustering Large Applications Based on Randomized Search
CLDN3	-	Claudin 3
CLIQUE	-	Clustering in Quest
DAVID	-	Database for Annotation, Visualization, and Integrated Discovery
DB	-	Davies and Bouldin
DBSCAN	-	Density-Based Spatial Clustering of Applications with Noise
DEGs	-	Differentially Expressed Genes
DIANA	-	Divisive Analysis
DRW	-	Directed Random Walk
dwgLASSO	-	Differentially Weighted Graphical Least Absolute Shrinkage and Selection Operator
ECOC	-	Error Correcting Output Codes

eDRW+	-	An Enhanced Directed Random Walk
ERBB2	-	Receptor Tyrosine-Protein Kinase Erythroblastic Oncogene B-2
ESEA	-	Edge Set Enrichment Analysis
expO	-	Expression Project for Oncology
FCM	-	Fuzzy C Means
FCS	-	Functional Class Scoring
GANPA	-	Network-based Gene Weights
GAT	-	Gene-Set Activity Toolbox
GCHL	-	Grid-Clustering Technique for High-Dimensional and Large Spatial Databases
GCS	-	Gene Expression Change Score
GEO	-	Gene Expression Omnibus
GO	-	Gene Ontology
GRIDEN	-	Grid-based and Density-based Spatial Clustering
GRPDBCAN	-	Grid-based DBSCAN Technique with Referential Parameters
GWAS	-	Genome-Wide Association Studies
HCI	-	High-order Correlation Integration
HPRD	-	Human Protein Reference Database
ID	-	Identifier
K-NN	-	K-Nearest Neighbours
KEGG	-	Kyoto Encyclopaedia of Genes and Genomes
LCC	-	Large Cell Carcinoma
MCWalk	-	Monte Carlo Simulations with Directed Random Walk

MSE	-	Mean Square Error
NCFS	-	Negatively Correlated Feature Sets
NCFS-CORG	-	Negatively Correlated Feature Sets with Condition-Responsive Genes
NCFS-i	-	Negatively Correlated Feature Sets with Ideal Markers
NSCLC	-	Non-Small Cell Lung Cancer
OMIM	-	Online Mendelian Inheritance in Man
ORA	-	Overrepresentation Analysis
OVA	-	One-Versus-All
OVO	-	One-Versus-One
PAM	-	Partitioning Around Medoids
PARADIGM	-	Pathway Recognition Algorithm using Data Integration on Genomic Models
PCA	-	Principal Component Analysis
PCC	-	Pearson Correlation Coefficients
PerPAS	-	Personalized Pathway Alteration Analysis
PMIDs	-	PubMed Identifiers
RMA	-	Robust Multichip Average
ROC	-	Receiver Operating Characteristic
RRFE	-	Reweighted Recursive Feature Elimination
SAM	-	Significance Analysis of Microarray
SCC	-	Squamous Cell Carcinoma
skeDRW+	-	Integration of K-Means Clustering and Average Silhouette Method into Enhanced Directed Random Walk
SOM	-	Self-Organizing Maps
SPIA	-	Signalling Pathway Impact Analysis

STING	-	Statistical Information Grid
SVM	-	Support Vector Machine
SVM-RFE	-	Support Vector Machine-Recursive Feature Elimination
TP53	-	Tumor protein p53
TPEA	-	Topology-Based Pathway Enrichment Analysis
UTM	-	Universiti Teknologi Malaysia
VIF	-	Variance Inflation Factor
Weighted-SAMGSR	-	Weighted-Significance Analysis of Microarray-Gene Set Reduction
WEKA	-	Waikato Environment for Knowledge Analysis
wgLASSO	-	Weighted Graphical Least Absolute Shrinkage and Selection Operator
WHO	-	World Health Organization

LIST OF SYMBOLS

$absolute\ F_{test}$	-	Absolute Values of F-test statistic
C	-	Row-Normalized Adjacency Matrix of the Selected Gene Clusters with Silhouette Width Values in the range of 0 and 1
$F_{test}(g_i)$	-	F-test statistics of Gene i from One-Way ANOVA on Expression Values between Multiple Classes of Samples
$F_{test}(P)$	-	F-test statistics of P from One-Way ANOVA on Pathway Activities between Multiple Classes of Samples
$gene$	-	Gene Expression Values for Gene over All the Samples
$maximum\ F_{test}$	-	Maximum Values of F-test statistic
$\bar{X}$	-	Mean of Gene Expression Values for Gene
$minimum\ F_{test}$	-	Minimum Values of F-test statistic
M^T	-	Row-Normalized Adjacency Matrix of a Directed Graph
P_j	-	Pathway Activity in Row j
r	-	Restart Probability
$sgn()$	-	Sign Function
S	-	Standard Deviation of Gene Expression Values for Gene

W_0	-	Initial Weight of Gene
W_t	-	A vector that Holds the Probability at the Specific Node at Time Step t
W_∞	-	Weight of Gene
$\lvert X'X \rvert$	-	Determinant
$z(g_i)$	-	Normalized Gene Expression Values for Gene over All the Samples

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Cancer is caused by cells which grow uncontrollably (Makropoulou, 2016). This disease is associated with abnormal alterations that lead to the dysregulation of the cellular system (Vaske *et al.*, 2010). According to the report of World Health Organization (WHO) in 2012, cancer contributes to approximately 14 million new cases and 8.2 million deaths. Bioinformatics develops computational methods to understand the molecular basis of disease (Napier and Limogiannis, 2016). The improved understanding of molecular biology and cellular biology has led to new cancer treatments since Richard Nixon (United States President) declared the "War on Cancer" in 1971. The cancer death rate was then declined by five percent between 1950 and 2005. Accurate classification of diseases and treatment responses is helpful in clinical and cancer research (Vaske *et al.*, 2010; Liu *et al.*, 2013a; Mohapatra *et al.*, 2016). The classification can identify groups of patients who share similar clinical features (characteristics) for the identification and implementation of suitable treatment (Macher and Crocq, 2004). Integrating pathway and topology information into microarray analysis can reduce the complexity of analysis from thousands of genes to a few hundreds of pathways (AlAjlan and Badr, 2015). This analysis is also aimed to identify more robust cancer markers to the disease of interest (Shi *et al.*, 2018).

1.2 Problem Background

Figure 1.1 presented an overview of the computational method to use in cancer classification. The common problem of cancer classification is the nature of cancer datasets, which have thousands of genes and characterized by small sample sizes based on different conditions (Su *et al.*, 2010; Jia *et al.*, 2011). In the literature, the use of microarray is different from macroarray, especially in term of probe density. Microarray contained a higher number of probes and such higher density of probes than macroarray (Vrana *et al.*, 2003). Macroarray was unique because it used radioactive target labelling for detection (Gammill and Lee, 2008). Since each picked clone must be sequenced to identify its identity, macroarray poorly annotated for potential novel genes. In the field of bioinformatics, microarray analysis is useful to measure the change of gene expression level in cancer datasets (Grewal and Das, 2013; Rajkumar *et al.*, 2013; Chandra and Babu, 2014). It is insufficient to use gene expression data only for microarray analysis, such as principal component analysis (PCA) in combination with agglomerative hierarchical clustering (AHC), mean-centering and magnitude normalization (Yasrebi *et al.*, 2009; Karn *et al.*, 2010).

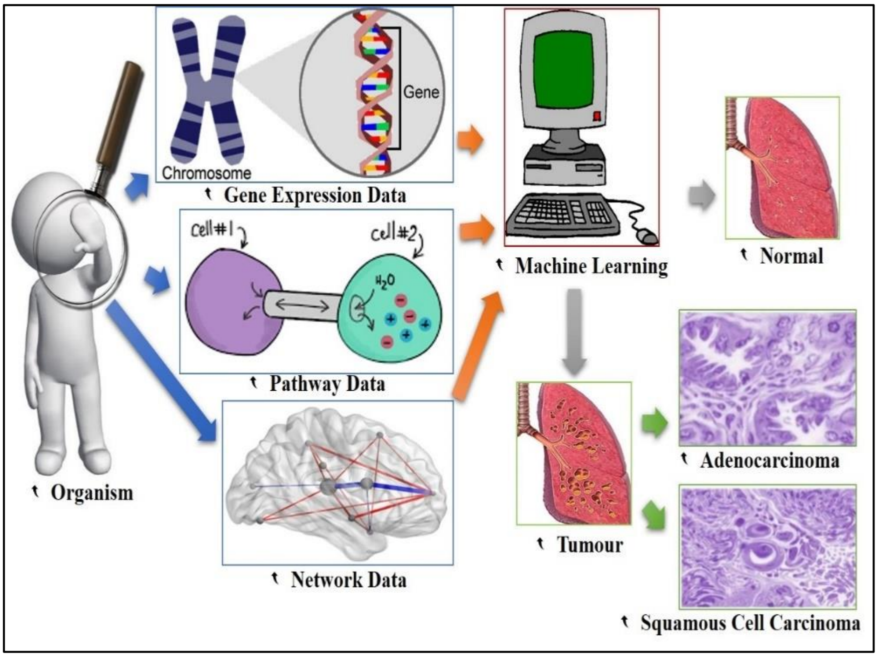


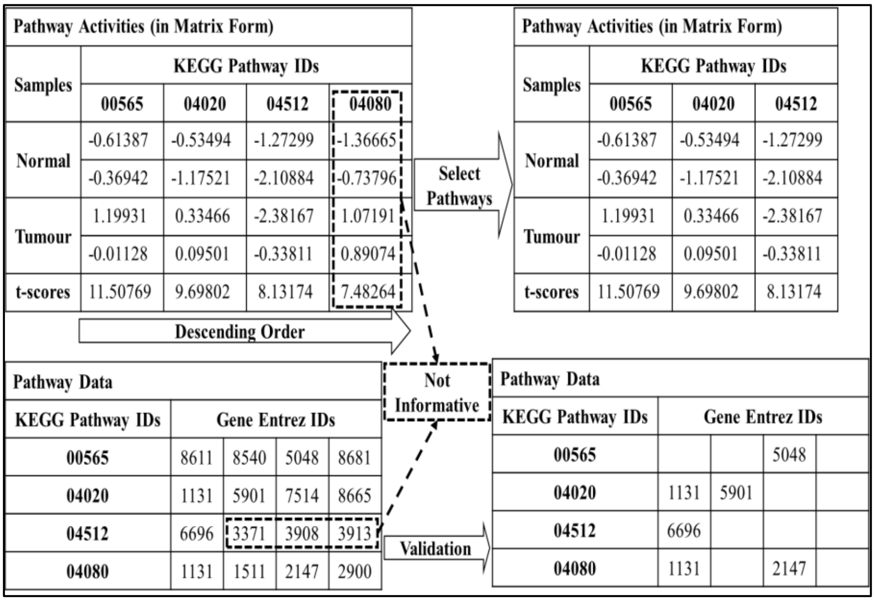
Figure 1.1 An overview of the pathway topology-based microarray analysis.

Pathway topology-based microarray analysis (e.g., Directed Random Walk [DRW]) is one of the categories for pathway-based microarray analysis, which can map genes on the precompiled pathways to visualize the whole chain of events in gene expression data (Grewal and Das, 2013). Since pathway topology-based microarray analysis can interpret pathways from the gene expression levels, pathway marker was more reliable than gene marker. The pathways were functionally related to the specified member genes with similar molecular mechanisms based on cancer subtypes (Zhao *et al.*, 2011; Hung and Chiu, 2017). Since tumour profiling of patients annotated in clinical practice, cancer markers were potentially further studied for new drug development and decision making in oncology to increase cancer survival (Wang *et al.*, 2015). DRW used weighting strategy to create weights for each gene in the directed graph based on the pathway knowledge to infer a higher reproducibility power of pathway activity (Liu *et al*., 2013a; Tian *et al.*, 2016). This method can reduce the effect of noise measurements and a correlation between genes in the same pathway (Su *et al.*, 2010). Besides, restart probability ($r = 0.7$) was the only parameter of DRW to characterize the level of strongly connected genes (e.g., a neighbourhood can be influenced by a seed gene to its neighbour gene) in the directed graph (Liu *et al*., 2013a; Wang *et al.*, 2017). The process of DRW with restart probability was iterated until all genes were visited.

In general, pathway activity is the formation of gene expression data and pathway data (with directed graph) by pathway topology-based microarray analysis. The analysis of the directed graph can reflect the functional robustness of topology in vital biochemical processes (Zhao *et al*., 2011; Roy *et al*., 2019). All the pathways used in the research were converted to a directed graph using *SubpathwayMiner* in R software package and its information was retrieved from the pathway database (e.g., Kyoto Encyclopaedia of Genes and Genomes [KEGG]) (Liu *et al*., 2013a; Dimitrakopoulos and Beerenwinkel, 2017). The topological information of the directed graph included types of interaction between two genes (direction of the edges), the weight of genes, and such position of genes. The interaction types between two genes showed how the two genes interacted and regulated each other in the processes of inhibition or activation.

The most common cancer deaths are caused in lung, breast, liver, colon, oesophagus, and stomach. Breast cancer is the most common cancer in women across every single ethnic group in Malaysia (Beshir and Hanipah, 2012; Nies *et al.*, 2017b). The breast cancer molecular subtypes are luminal A, luminal B, basal, ERBB2, and normal. The main subtypes of lung cancer are adenocarcinoma, large cell carcinoma, and squamous cell carcinoma. Some existing methods of pathway-based microarray analysis are restricted to classify the datasets between normal and tumour samples with the use of t-test, such as negatively correlated feature sets with ideal markers (NCFS-i), negatively correlated feature sets with condition-responsive genes (NCFS-CORG), and DRW (Chan *et al.*, 2011; Chandra and Gupta, 2011; Sootanan *et al.*, 2011; Liu *et al.*, 2013a; Yang *et al.*, 2014; Phongwattana *et al.*, 2015; Ross and Willson, 2017). Besides, some methods modify t-test and ANOVA to deal with multiclass issues, such as weighted-significance analysis of microarray-gene set reduction (Weighted-SAMGSR), negatively correlated feature sets (NCFS), gene-set activity toolbox, and ANOVA-based feature set (AFS) (Chen *et al.*, 2005; Engchuan and Chan, 2012, 2015; Engchuan *et al.*, 2016; Kar *et al.*, 2016; Tian *et al.*, 2016; Ortiz-Ramón *et al.*, 2018). Multiclass classification methods can be divided into two types. First, this involves extending the binary classification to deal with the multiclass problems directly (Li *et al.*, 2004; Ferdowsi *et al.*, 2014). Another type involves decomposing multiclass issues into binary problems. One-versus-one and one-versus-the-rest are common strategies for dealing with multiclass problems, but some are not extensible to multiclass approaches (Gu *et al.*, 2014; Ferdowsi *et al.*, 2014). To date, recent medical studies reported the necessity to diagnose more than two classes of disease (Engchuan and Chan, 2012, 2015; Yang and Naiman, 2014; Yang *et al.*, 2014). Clinical experiments can produce multiclass gene expression data in the detection of tumours based on their stage, grade, survival time, and drug sensitivity that are further studied for cancer treatments (Yang and Naiman, 2014; Wang *et al.*, 2015). For example, stages of such disease depend on the thickness of tumour at the time of surgical treatment.

Several studies in pathway-based microarray analysis do not select pathways, including DRW. Since pathways were commonly curated from the literature, non-informative genes can be included and affect the accuracy of the methods (Evangeline *et al.*, 2013; Zhe *et al.*, 2013; Creixell *et al.*, 2015; Li *et al.*, 2017). If a gene (e.g., tumor protein p53) is chosen, all the pathways (e.g., cell cycle and MAPK signaling pathway) consist of such gene will also be selected. Figure 1.2 illustrated the presence of non-informative genes in a pathway. Pathway selection can reduce the dimension and select informative pathways in all the examples (Zhe *et al.*, 2013; Gu *et al.*, 2014). With cases of existing methods performed pathway selection using t-test and Fisher-test, such as redundancy removable pathway-based feature selection method and the network and node selection approach.



Pathway Activities (in Matrix Form)

Samples	KEGG Pathway IDs			
	00565	04020	04512	04080
Normal	-0.61387	-0.53494	-1.27299	-1.36665
	-0.36942	-1.17521	-2.10884	-0.73796
Tumour	1.19931	0.33466	-2.38167	1.07191
	-0.01128	0.09501	-0.33811	0.89074
t-scores	11.50769	9.69802	8.13174	7.48264

Pathway Activities (in Matrix Form)

Samples	KEGG Pathway IDs		
	00565	04020	04512
Normal	-0.61387	-0.53494	-1.27299
	-0.36942	-1.17521	-2.10884
Tumour	1.19931	0.33466	-2.38167
	-0.01128	0.09501	-0.33811
t-scores	11.50769	9.69802	8.13174

Pathway Data

KEGG Pathway IDs	Gene Entrez IDs			
00565	8611	8540	5048	8681
04020	1131	5901	7514	8665
04512	6696	3371	3908	3913
04080	1131	1511	2147	2900

Pathway Data

KEGG Pathway IDs	Gene Entrez IDs			
00565			5048	
04020	1131	5901		
04512	6696			
04080	1131		2147	

Figure 1.2 The presence of non-informative genes in a pathway.

In literature, random walk used the theory of Markov chain to rank genes from high to low probabilities, but it extracted local information from a large graph without knowledge of the whole graph data (Liu *et al.*, 2013a, 2017b; Liu *et al.*, 2015b; Zhang *et al.*, 2016; Dimitrakopoulos and Beerenwinkel, 2017; Wang and Liu, 2018; Peng *et al.*, 2019). Hence, a large directed graph can include non-informative genes, which can result in low accuracy of the methods (Evangeline *et al.*, 2013; Peng *et al.*, 2019). Besides, DRW used the theory of random walk to identify the genes having similar structural properties of networks (Re and Valentini, 2012; Petrochilos *et al.*, 2013). However, most methods (including DRW) ignored the interaction between multiple genes in a directed graph and the genes with missing edges (Madhukar *et al.*, 2015; Liu *et al.*, 2017a). Figure 1.3 illustrated the common neighbour and non-informative genes in a directed graph. A gene was important (e.g., gene Entrez ID 5901) if it interacted with many other genes (Zhu *et al.*, 2018).

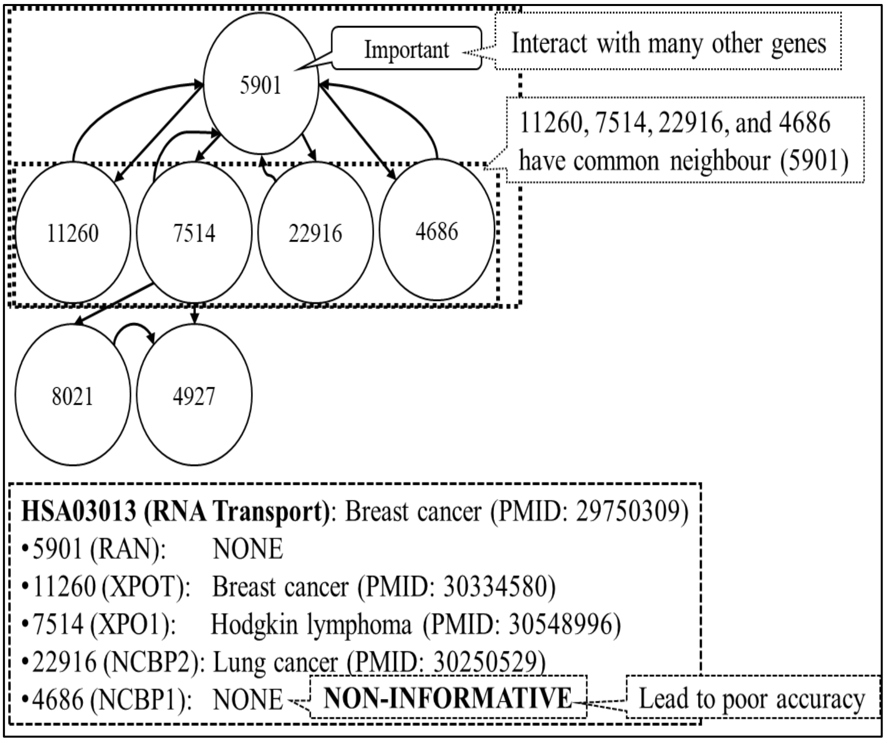

5901
Important
Interact with many other genes
11260, 7514, 22916, and 4686 have common neighbour (5901)
11260
7514
22916
4686
8021
4927
HSA03013 (RNA Transport): Breast cancer (PMID: 29750309)
•5901 (RAN): NONE
•11260 (XPOT): Breast cancer (PMID: 30334580)
•7514 (XPO1): Hodgkin lymphoma (PMID: 30548996)
•22916 (NCBP2): Lung cancer (PMID: 30250529)
•4686 (NCBP1): NONE
NON-INFORMATIVE
Lead to poor accuracy

Figure 1.3 The presence of common neighbour and non-informative genes in a directed graph.

Several previous studies have noted the importance of clustering to identify co-expressed genes in a cluster and inactive genes in another cluster (Mehmood *et al.*, 2018; Chandra and Tripathi, 2019). Clustering can also discover the fundamental hidden structure of biomedical data and identify cancer subtypes that used for diagnosis and treatments. DRW is also one of the density-based clustering techniques, but it has a high runtime analysis to detect clusters (Deng *et al.*, 2018b). Detection of clusters using partitioning clustering has low time complexity and high computing efficiency, which can solve the issue above (Xu and Tian, 2015). Researchers focused on partitioning clustering techniques (e.g., k-means clustering) by assuming the number of clusters beforehand, which can lead to the poor quality of clusters (Bajo *et al.*, 2010; Wang *et al.*, 2018a; Majhi and Biswal, 2019).

1.3 Problem Statements

Pathway topology-based microarray analysis used pathway data, directed graph, and gene expression data to identify pathway and gene markers in cancer classification. However, most existing techniques analyse the datasets by grouping different cancer subtypes into disease sample only. All the pathways consisted of the specified gene were selected and considered these pathways equally. Several current methods ignore the genes with missing edges and the interaction between multiple genes. All the issues can lead to low accuracy and large-scale variation in weight vectors. Partitioning clustering techniques are useful to detect clusters, but it can lead to poor quality of clustering by assuming to initialize the number of potential clusters beforehand.

The main research question of this research is:

How to identify pathway and gene markers for multiclass cancer expression data in order to improve the use of weight strategy in pathway topology-based microarray analysis?

Thus, the following issues will be considered to solve the problem:

- *How to identify pathway markers between multiple classes of samples in order to improve the weight of genes?*

- *How to identify pathway markers from all the pathways and increase the accuracy of the method for multiclass cancer expression data?*

- *How to identify the number of potential clusters needed to initialize for k-means clustering technique in order to improve the quality of clustering?*

- *How to integrate k-means clustering and average silhouette method into the method in the directed graph for identifying gene markers for multiclass cancer expression data?*

1.4 Research Goal

The goal of the research is to propose enhanced directed random walk with improved use of weight strategy in topology-based microarray analysis and consideration of the interaction between genes for identification of pathway and gene markers from multiclass cancer expression data.

1.5 Research Objectives

The objectives of the research are:

- To propose an enhanced directed random walk method (eDRW+) for identification of pathway markers from multiclass cancer expression data to improve the use of weight strategy and pathway selection based on the greatest reproducibility power.
- To propose skeDRW+ based on the integration of k-means clustering and average silhouette method into eDRW+ for identification of gene markers from multiclass cancer expression data in order to improve the quality of clustering.
- To biologically validate pathway and gene markers using PubMed text data mining and functional enrichment analysis in pathway data, directed graph, and gene expression data.

1.6 Research Scopes

This research focuses on the identification of cancer markers and emphasizes the issues of pathway topology-based microarray analysis. This research also aims to improve the weight of genes and improve the quality of clustering for identifying similar biological functions of genes. Figure 1.4 illustrated the flow of the research from bioinformatics to the discovery of cancer markers.

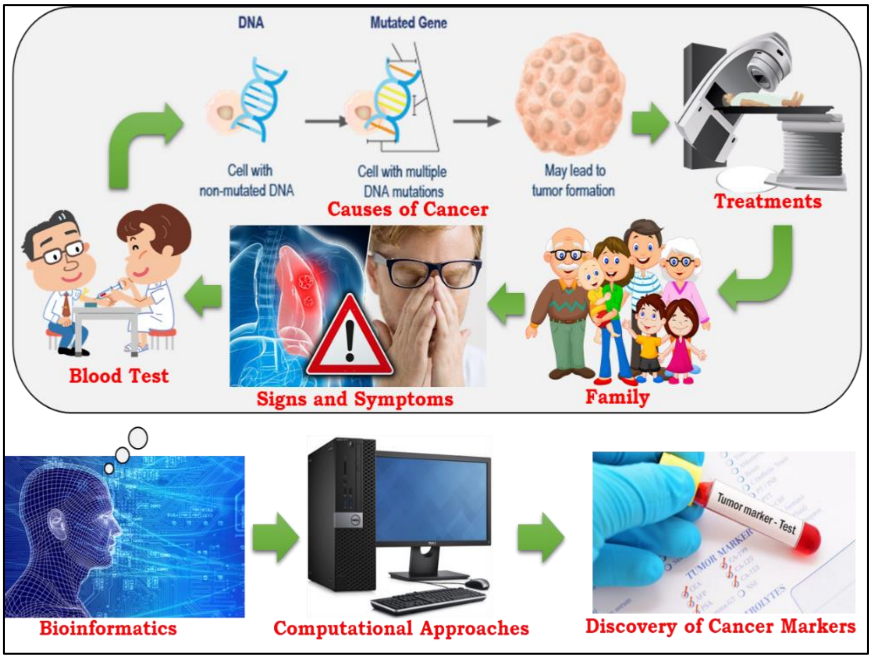


Figure 1.4 The flow of the research from bioinformatics to the discovery of cancer markers.

The following points are the research scopes:

- According to the research focus, three components constitute the scopes of the research. The research investigates *directed random walk method (DRW)* [WHAT] as pathway topology-based method in identifying pathway and gene markers for *multiclass cancer expression data* [WHERE] in order to improve *survival and quality of life* [WHY]. The cancer markers can identify drug targets and look for cancer subtypes with clinically distinct outcomes.

- This research uses gene expression data (lung, breast, and skin cancers), pathway data (metabolic and non-metabolic pathways), and directed graph.

- The development of the proposed methods is implemented in the R platform with version 3.3.3.

- The performance of this research is measured in a stratified ten-fold cross-validation, which was mostly used in previous works. The experimental results are compared in terms of area under the receiver operating characteristics curve (AUC) and accuracy (%) to justify the performance improvement.

- The identified pathways and genes are biologically validated using PubMed text data mining and functional enrichment analysis to show the relationship between pathways, genes, and cancers.

1.7 Research Significances

This research is considered significant as it tends to identify pathway and gene markers for multiclass cancer expression data using pathway topology-based microarray analysis. This method used multiple data types to infer a greater reproducibility power of pathway activity with higher classification accuracy. There is a need to classify the datasets into multiple classes of samples, which can deal with grouping different cancer subtypes into disease sample only. The use of pathway data can help to study molecular mechanisms based on cancer subtypes. Pathway selection can identify more pathway markers, although the non-informative genes included in the pathways. To increase the efficiency of identifying gene markers and improve the weight of genes in the directed graph, partitioning clustering and optimization techniques integrated into pathway topology-based microarray analysis can reduce the variation in weight vectors and improve its quality during initialization of the intended cluster number. Furthermore, the identified pathway and gene markers can be further used in new drug development and clinical implications for cancers. It can also help patients having early detection and diagnosis.

1.8 Thesis Organization

This thesis is organized into six chapters. The flow of the following chapters is presented as follows. Chapter 2 aims to describe some basic knowledge related to this research. This chapter also includes reviewing some preliminary collections of present works done in previous studies related to this research area. Chapter 3 aims to discuss the details of the research methodology employed in this research. The research framework is explained in this chapter to achieve the goal and objectives of the research. This chapter also includes input data, a summary of the proposed methods, software and hardware requirements used in this research. Performance measurements are also explained in this chapter to evaluate and compare among the methods. Chapter 4 aims to present the proposed methods, eDRW+ and skeDRW+, in identifying pathway and gene markers for multiclass cancer expression data. Chapter 5 aims to present and discuss the experimental results generated by eDRW+ and skeDRW+. Chapter 6 aims to conclude the findings, contribution, and suggestions for future works in this research.

REFERENCES

Aarthi, P., and Gothai, E. (2014). Improving Class Separability for Microarray datasets using Genetic Algorithm with KLD Measure. *Int. J. Eng. Sci. Innov. Technol*, *3*(2), 514-521.

Abraham, A., Das, S., and Roy, S. (2008). Swarm intelligence algorithms for data clustering. In *Soft computing for knowledge discovery and data mining* (pp. 279-313). Springer, Boston, MA.

Acharya, S., Saha, S., and Sahoo, P. (2019). Bi-clustering of microarray data using a symmetry-based multi-objective optimization framework. *Soft Computing*, *23*(14), 5693-5714.

Adeleye, Y., Andersen, M., Clewell, R., Davies, M., Dent, M., Edwards, S., Fowler, P., Malcomber, S., Nicol, B., Scott, A. and Scott, S. (2015). Implementing Toxicity Testing in the 21st Century (TT21C): Making safety decisions using toxicity pathways, and progress in a prototype risk assessment. *Toxicology*, *332*, 102-111.

AlAjlan, A., and Badr, G. (2015, July). Data Mining in Pathway Analysis for Gene Expression. In *Industrial Conference on Data Mining* (pp. 69-77). Springer, Cham.

Aly, M. (2005). *Survey on multiclass classification methods*. Neural networks, 19, pp. 1–9.

An, J., Kim, K., Chae, H., and Kim, S. (2014). DegPack: a web package using a non-parametric and information theoretic algorithm to identify differentially expressed genes in multiclass RNA-seq samples. *Methods*, *69*(3), 306-314.

Azzawi, H., Hou, J., Alanni, R., Xiang, Y., Abdu-Aljabar, R., and Azzawi, A. (2017, November). Multiclass lung cancer diagnosis by gene expression programming and microarray datasets. In *International Conference on Advanced Data Mining and Applications* (pp. 541-553). Springer, Cham.

Bäck, T., Fogel, D. B., and Michalewicz, Z. (Eds.). (2018). *Evolutionary computation 1: Basic algorithms and operators*. CRC press.

Bäck, T., Rudolph, G., and Schwefel, H. P. (1993, February). Evolutionary programming and evolution strategies: Similarities and differences. In *In Proceedings of the Second Annual Conference on Evolutionary Programming*.

Bajo, J., De Paz, J. F., Rodríguez, S., and González, A. (2011). A new clustering algorithm applying a hierarchical method neural network. *Logic Journal of IGPL*, *19*(2), 304-314.

Bandyopadhyay, S., Saha, S., Maulik, U., and Deb, K. (2008). A simulated annealing-based multiobjective optimization algorithm: AMOSA. *IEEE transactions on evolutionary computation*, *12*(3), 269-283.

Bao, Z., Zhu, Y., Ge, Q., Gu, W., Dong, X., and Bai, Y. (2019). gwSPIA: Improved Signaling Pathway Impact Analysis With Gene Weights. *IEEE Access*, *7*, 69172-69183.

Baranzini, S. E., Galwey, N. W., Wang, J., Khankhanian, P., Lindberg, R., Pelletier, D., Wu, W., Uitdehaag, B.M., Kappos, L., GeneMSA Consortium and Polman, C. H. (2009). Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Human molecular genetics*, *18*(11), 2078-2090.

Bassani, H. F., and Araujo, A. F. (2014). Dimension selective self-organizing maps with time-varying structure for subspace and projected clustering. *IEEE transactions on neural networks and learning systems*, *26*(3), 458-471.

Bayerlová, M., Jung, K., Kramer, F., Klemm, F., Bleckmann, A., and Beißbarth, T. (2015). Comparative study on gene set and pathway topology-based enrichment methods. *BMC bioinformatics*, *16*(1), 334.

Ben-Hamo, R., Gidoni, M., and Efroni, S. (2014). PhenoNet: identification of key networks associated with disease phenotype. *Bioinformatics*, *30*(17), 2399-2405.

Bénichou, O., Cazabat, A. M., Moreau, M., and Oshanin, G. (1999). Directed random walk in adsorbed monolayer. *Physica A: Statistical Mechanics and its Applications*, *272*(1-2), 56-86.

Bernhardson, C. S. (1975). 375: Type I error rates when multiple comparison procedures follow a significant F test of ANOVA. *Biometrics*, 229-232.

Besaw, M. E. (2013). Protein lounge. *Journal of the Medical Library Association: JMLA*, *101*(2), 164.

Beshir, S. A., and Hanipah, M. A. (2012). Knowledge, perception, practice and barriers of breast cancer health promotion activities among community pharmacists in two Districts of Selangor state, Malaysia. *Asian Pacific Journal of Cancer Prevention*, *13*(9), 4427-4430.

Bholowalia, P., and Kumar, A. (2014). EBK-means: A clustering technique based on elbow method and k-means in WSN. *International Journal of Computer Applications*, *105*(9).

Billmann, M., Chaudhary, V., ElMaghraby, M. F., Fischer, B., and

Boutros, M. (2018). Widespread rewiring of genetic networks upon cancer signaling pathway activation. *Cell systems*, *6*(1), 52-64.

Brazma, A., and Vilo, J. (2000). Gene expression data analysis. *FEBS letters*, *480*(1), 17-24.

Breitkreutz, D., Hlatky, L., Rietman, E., and Tuszynski, J. A. (2012). Molecular signaling network complexity is correlated with cancer patient survivability. *Proceedings of the National Academy of Sciences*, *109*(23), 9209-9212.

Brown, I., and Mues, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, *39*(3), 3446-3453.

Bryant, A., and Cios, K. (2018). RNN-DBSCAN: A density-based clustering algorithm using reverse nearest neighbor density estimates. *IEEE Transactions on Knowledge and Data Engineering*, *30*(6), 1109-1121.

Cao, Y., Lu, Y., Pan, X. and Sun, N. (2019). An improved global best guided artificial bee colony algorithm for continuous optimization problems. *Cluster computing*, *22*(2), 3011-3019.

Carneiro, M. G., Cheng, R., Zhao, L. and Jin, Y. (2019). Particle swarm optimization for network-based data classification. *Neural Networks*, *110*, 243-255.

Carson, M. B. and Lu, H. (2015). Network-based prediction and knowledge mining of disease genes. *BMC medical genomics*, *8*(2), S9.

Chan, J. H., Sootanan, P. and Larpeampaisarl, P. (2011, July). Feature selection of pathway markers for microarray-based disease classification using negatively correlated feature sets. In *The*

2011 International Joint Conference on Neural Networks (pp. 3293-3299). IEEE.

Chandra, B. and Babu, K. N. (2014). Classification of gene expression data using spiking wavelet radial basis neural network. *Expert systems with applications*, *41*(4), 1326-1330.

Chandra, B. and Gupta, M. (2011). An efficient statistical feature selection approach for classification of gene expression data. *Journal of biomedical informatics*, *44*(4), 529-535.

Chandra, G. and Tripathi, S. (2019). A Column-Wise Distance-Based Approach for Clustering of Gene Expression Data with Detection of Functionally Inactive Genes and Noise. In *Advances in Intelligent Computing* (pp. 125-149). Springer, Singapore.

Che, J., Yue, D., Zhang, B., Zhang, H., Huo, Y., Gao, L., Zhen, H., Yang, Y. and Cao, B. (2018). Claudin-3 inhibits lung squamous cell carcinoma cell epithelial-mesenchymal transition and invasion via suppression of the Wnt/β-catenin signaling pathway. *International journal of medical sciences*, *15*(4), 339.

Chehouri, A., Younes, R., Khoder, J., Perron, J. and Ilinca, A. (2017). A selection process for genetic algorithm using clustering analysis. *Algorithms*, *10*(4), 123.

Chen, D., Liu, Z., Ma, X. and Hua, D. (2005). Selecting genes by test statistics. *BioMed Research International*, *2005*(2), 132-138.

Chen, X., Li, J., Gray, W. H., Lehmann, B. D., Bauer, J. A., Shyr, Y. and Pietenpol, J. A. (2012). TNBCtype: a subtyping tool for triple-negative breast cancer. *Cancer informatics*, *11*, CIN-S9983.

Chen, Y., Tang, S., Bouguila, N., Wang, C., Du, J. and Li, H. (2018). A fast clustering algorithm based on pruning unnecessary distance computations in DBSCAN for high-dimensional data. *Pattern Recognition*, *83,* 375-387.

Cheng, C. and Bao, C. (2017, February). A kernelized fuzzy C-means clustering algorithm based on glowworm swarm optimization algorithm. In *Proceedings of the 9th International Conference on Computer and Automation Engineering* (pp. 78-82).

Cheng, W., Wang, W. and Batista, S. (2018). Grid-Based Clustering. *Data Clustering,* pp. 128–148, London, United Kingdom: Chapman and Hall, CRC.

Civicioglu, P. and Besdok, E. (2013). A conceptual comparison of the Cuckoo-search, particle swarm optimization, differential evolution and artificial bee colony algorithms. *Artificial Intelligence Review*, *39(4),* pp. 315–346.

Clarke, P. A., te Poele, R., Wooster, R. and Workman, P. (2001). Gene expression microarray analysis in cancer biology, pharmacology, and drug development: progress and potential. *Biochemical pharmacology*, *62(10),* pp. 1311–1336.

Creixell, P., Reimand, J., Haider, S., Wu, G., Shibata, T., Vazquez, M., Mustonen, V., Gonzalez-Perez, A., Pearson, J., Sander, C. and Raphael, B. J. (2015). Pathway and network analysis of cancer genomes. *Nature methods*, *12(7),* 615.

Das, D., Pratihar, D. K., Roy, G. G. and Pal, A. R. (2018). Phenomenological model-based study on electron beam welding process, and input-output modeling using neural networks trained by back-propagation algorithm, genetic algorithms, particle swarm optimization algorithm and bat algorithm. *Applied Intelligence, 48(9)*, pp. 2698–2718.

Datta, Susmita and Datta, Somnath (2006). Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics, 7(1)*, p. 397.

Davidson, I. and Ravi, S. S. (2005, October). Agglomerative Hierarchical Clustering with Constraints: Theoretical and Empirical Results. In *European Conference on Principles of Data Mining and Knowledge Discovery* (pp. 59-70). Springer, Berlin, Heidelberg.

de Barros Franco, D. G. and Steiner, M. T. A. (2018). Clustering of solar energy facilities using a hybrid fuzzy c-means algorithm initialized by metaheuristics. *Journal of cleaner production, 191*, pp. 445–457.

Dembele, D. and Kastner, P. (2003). Fuzzy C-means method for clustering microarray data. *Bioinformatics, 19(8)*, pp. 973–980.

Deng, C., Song, J., Sun, R., Cai, S. and Shi, Y. (2018a). GRIDEN: An effective grid-based and density-based spatial clustering algorithm to support parallel computing. *Pattern Recognition Letters, 109*, pp. 81–88.

Deng, C., Song, J., Sun, R., Cai, S. and Shi, Y. (2018b). Gridwave: a grid-based clustering algorithm for market transaction data based on spatial-temporal density-waves and synchronization. *Multimedia Tools and Applications, 77(22)*, pp. 29623–29637.

Dimitrakopoulos, C. M. and Beerenwinkel, N. (2017). Computational approaches for the identification of cancer genes and pathways. *Wiley Interdisciplinary Reviews: Systems Biology and Medicine, 9(1),* p. e1364.

Ding, F., Wang, J., Ge, J. and Li, W. (2018). Anomaly detection in large-scale trajectories using hybrid grid-based hierarchical clustering. *International Journal of Robotics and Automation, 33(5),* pp. 474–480.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P. and Yasui, Y. (2008). Gene-set analysis and reduction. *Briefings in bioinformatics, 10(1),* pp. 24–34.

Dong, X., Hao, Y., Wang, X. and Tian, W. (2016). LEGO: a novel method for gene set over-representation analysis by incorporating network-based gene weights. *Scientific reports, 6,* p. 18871.

Doungpan, N., Engchuan, W., Chan, J. H. and Meechai, A. (2016). GSNFS: Gene subnetwork biomarker identification of lung cancer expression data. *BMC Medical Genomics, 9(S3),* p. 70.

Doungpan, N., Engchuan, W., Meechai, A. and Chan, J. H. (2015, July). Clustering-based gene-subnetwork biomarker identification using gene expression data. In *2015 International Joint Conference on Neural Networks (IJCNN) (pp. 1-7). IEEE.*

Duarte, E. and Wainer, J. (2017). Empirical comparison of cross-validation and internal metrics for tuning SVM hyperparameters. *Pattern Recognition Letters, 88,* pp. 6–11.

Duval, B., Hao, J. K. and Hernandez Hernandez, J. C. (2009, July). A memetic algorithm for gene selection and molecular classification of cancer. In *Proceedings of the 11th Annual* *conference on Genetic and evolutionary computation* (pp. 201-208).

Efroni, S., Schaefer, C. F. and Buetow, K. H. (2007). Identification of Key Processes Underlying Cancer Phenotypes Using Biologic Pathway Analysis. *PLoS ONE, 2(5),* e425.

Eiben, A. E. and Smith, J. (2015). From evolutionary computation to the evolution of things. *Nature, 521(7553),* p. 476.

Enerly, E., Steinfeld, I., Kleivi, K., Leivonen, S.-K., Aure, M. R., Russnes, H. G., Rønneberg, J. A., Johnsen, H., Navon, R., Rødland, E., Mäkelä, R., Naume, B., Perälä, M., Kallioniemi, O., Kristensen, V. N., Yakhini, Z. and Børresen-Dale, A.-L. (2011). miRNA-mRNA Integrated Analysis Reveals Roles for miRNAs in Primary Breast Tumors. *PLoS ONE, 6(2),* e16915.

Engchuan, W. and Chan, J. H. (2012, November). Pathway-Based Multi-class Classification of Lung Cancer. In *International Conference on Neural Information Processing* (pp. 697-702). Springer, Berlin, Heidelberg.

Engchuan, W. and Chan, J. H. (2013). Apriori gene set-based microarray analysis for disease classification using unlabeled data. *Procedia Computer Science, 23,* pp. 137–145.

Engchuan, W. and Chan, J. H. (2015). Pathway activity transformation for multi-class classification of lung cancer datasets. *Neurocomputing, 165,* pp. 81–89.

Engchuan, W., Meechai, A., Tongsima, S., Doungpan, N. and Chan, J. H. (2016). Gene-set activity toolbox (GAT): A platform for microarray-based cancer diagnosis using an integrative gene-set analysis approach. *Journal of Bioinformatics and Computational Biology, 14(04),* p. 1650015.

Ester, M., Kriegel, H. P., Sander, J. and Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In *The Knowledge Discovery and Data Mining Conferences (KDD)* (Vol. 96, No. 34, pp. 226-231).

Evangeline, D. P., Sandhiya, C., Anandhakumar, P., Raj, G. D. and Rajendran, T. (2013, December). Feature subset selection for irrelevant data removal using Decision Tree Algorithm. In *2013 Fifth International Conference on Advanced Computing (ICoAC)* (pp. 268-274). IEEE.

Eyileten, C., Wicik, Z., De Rosa, S., Mirowska-Guzel, D., Soplinska, A., Indolfi, C., Jastrzebska-Kurkowska, I., Czlonkowska, A. and Postula, M. (2018). MicroRNAs as Diagnostic and Prognostic Biomarkers in Ischemic Stroke-A Comprehensive Review and Bioinformatic Analysis. *Cells, 7(12),* p. 249.

Farmer, P., Bonnefoi, H., Becette, V., Tubiana-Hulin, M., Fumoleau, P., Larsimont, D., MacGrogan, G., Bergh, J., Cameron, D., Goldstein, D., Duss, S., Nicoulaz, A.-L., Fiche, M., Brisken, C., Delorenzi, M. and Iggo, R. (2005). Identification of molecular apocrine breast tumours by microarray analysis. *Breast Cancer Research, 7(S2),* P2.11.

Ferdowsi, S., Voloshynovskiy, S., Gabryel, M. and Korytkowski, M. (2014, June). Multi-class Classification: A Coding Based Space Partitioning. In *International Conference on Artificial Intelligence and Soft Computing* (pp. 593-604). Springer, Cham.

Fister, I., Fister Jr, I., Yang, X. S. and Brest, J. (2013). A comprehensive review of firefly algorithms. *Swarm and*

Evolutionary Computation, 13, pp. 34–46.

Gammill, L. S. and Lee, V. M. (2008). Gene Discovery: Macroarrays and Microarrays. *Methods in cell biology, 87*, 297-312.

Gandomi, A. H., Yang, X.-S. and Alavi, A. H. (2013a). Cuckoo search algorithm: a metaheuristic approach to solve structural optimization problems. *Engineering with Computers, 29(1),* pp. 17–35.

Gandomi, A. H., Yang, X.-S., Alavi, A. H. and Talatahari, S. (2013b). Bat algorithm for constrained optimization tasks. *Neural Computing and Applications, 22(6),* pp. 1239–1255.

Gao, H., Jiang, J., She, L. and Fu, Y. (2010). A new agglomerative hierarchical clustering algorithm implementation based on the map reduce framework. *International Journal of Digital Content Technology and its Applications, 436(3),* pp. 95–100.

García, J., Crawford, B., Soto, R. and Astorga, G. (2019). A clustering algorithm applied to the binarization of Swarm intelligence continuous metaheuristics. *Swarm and Evolutionary Computation, 44*, pp. 646–664.

Garg, S. and Batra, S. (2018). Fuzzified cuckoo based clustering technique for network anomaly detection. *Computers and Electrical Engineering, 71,* pp. 798–817.

Garzón, J. A. C. and González, J. R. (2015). A gene selection approach based on clustering for classification tasks in colon cancer. *ADCAIJ: Advances in Distributed Computing and Artificial Intelligence Journal, 4(3),* pp. 1–10.

Geng, Y. A., Li, Q., Zheng, R., Zhuang, F., He, R. and Xiong, N. (2018). RECOME: A new density-based clustering algorithm using relative KNN kernel density. *Information Sciences, 436,* pp. 13–30.

Ghaedi, A. M., Ghaedi, M., Vafaei, A., Iravani, N., Keshavarz, M., Rad, M., Tyagi, I., Agarwal, S. and Gupta, V. K. (2015). Adsorption of copper (II) using modified activated carbon prepared from Pomegranate wood: optimization by bee algorithm and response surface. *Journal of Molecular Liquids, 206,* pp. 195–206.

Gibson, G. (2003). Microarray Analysis. *PLoS Biology, 1(1),* e15.

Gomez-Pilar, J., Poza, J., Bachiller, A., Gómez, C., Núñez, P., Lubeiro, A., Molina, V. and Hornero, R. (2018). Quantification of Graph Complexity Based on the Edge Weight Distribution Balance: Application to Brain Networks. *International Journal of Neural Systems, 28(01),* p. 1750032.

Grewal, R. K. and Das, S. (2013). Microarray data analysis: Gaining biological insights. *Journal of Biomedical Science and Engineering, 6(10),* p. 996.

Gu, J. L., Lu, Y., Liu, C. and Lu, H. (2014). Multiclass classification of sarcomas using pathway based feature selection method. *Journal of theoretical biology, 362,* pp. 3–8.

Gu, Z., Liu, J., Cao, K., Zhang, J. and Wang, J. (2012). Centrality-based pathway enrichment: a systematic approach for finding significant pathways dominated by key genes. *BMC Systems Biology, 6(1),* p. 56.

Gu, Z. and Wang, J. (2013). CePa: an R package for finding significant pathways weighted by multiple network centralities. *Bioinformatics, 29(5),* pp. 658–660.

Guebila, M. B., and Thiele, I. (2019). Predicting gastrointestinal drug effects using contextualized metabolic models. *PLoS* *computational biology*, *15*(6).

Guven, A. and Aytek, A. (2009). New Approach for Stage-Discharge Relationship: Gene-Expression Programming. *Journal of Hydrologic Engineering, 14(8),* pp. 812–820.

Haakensen, V. D., Steinfeld, I., Saldova, R., Shehni, A. A., Kifer, I., Naume, B., Rudd, P. M., Børresen-Dale, A. L. and Yakhini, Z. (2016). Serum N-glycan analysis in breast cancer patients–relation to tumour biology and clinical outcome. *Molecular oncology, 10(1),* pp. 59–72.

Halkidi, M. and Vazirgiannis, M. (2001). Clustering validity assessment: Finding the optimal partitioning of a data set. In *Proceedings 2001 IEEE International Conference on Data Mining* (pp. 187-194). IEEE.

Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P. and Witten, I. H. (2009). The WEKA data mining software: an update. *ACM SIGKDD explorations newsletter, 11(1),* pp. 10–18.

Han, J., Shi, X., Zhang, Y., Xu, Y., Jiang, Y., Zhang, C., Feng, L., Yang, H., Shang, D., Sun, Z. and Su, F. (2015). ESEA: discovering the dysregulated pathways based on edge set enrichment analysis. *Scientific reports, 5,* p. 13044.

Handhayani, T. and Hiryanto, L. (2015). Intelligent kernel k-means for clustering gene expression. *Procedia Computer Science, 59,* pp. 171–177.

Handl, J., Knowles, J. and Kell, D. B. (2005). Computational cluster validation in post-genomic data analysis. *Bioinformatics, 21(15),* pp. 3201–3212.

Hochreiter, S., Clevert, D. A. and Obermayer, K. (2006). A new summarization method for Affymetrix probe level data. *Bioinformatics, 22(8),* pp. 943–949.

Hu, J. and Pei, J. (2018). Subspace multi-clustering: a review. *Knowledge and Information Systems, 56(2),* pp. 257–284.

Huan, J., Wang, L., Xing, L., Qin, X., Feng, L., Pan, X. and Zhu, L. (2014). Insights into significant pathways and gene interaction networks underlying breast cancer cell line MCF-7 treated with 17β-estradiol (E2). *Gene, 533(1),* pp. 346–355.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2008). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research, 37(1),* pp. 1–13.

Huang, D. W., Sherman, B. T. and Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols, 4(1),* p. 44.

Hung, F. H. and Chiu, H. W. (2017). Cancer subtype prediction from a pathway-level perspective by using a support vector machine based on integrated gene expression and protein network. *Computer methods and programs in biomedicine, 141*, pp. 27–34.

Ibrahim, M. A.-H., Jassim, S., Cawthorne, M. A. and Langlands, K. (2012). A Topology-Based Score for Pathway Enrichment. *Journal of Computational Biology, 19(5),* pp. 563–573.

Ihnatova, I. and Budinska, E. (2015). ToPASeq: an R package for topology-based pathway analysis of microarray and RNA-Seq data. *BMC Bioinformatics, 16(1),* p. 350.

Imdadullah, M., Aslam, M. and Altaf, S. (2016). mctest: An R package for detection of collinearity among regressors. *The R Journal, 8(2),* pp. 499–509.

Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters, 31(8),* pp. 651–666.

Jang, I. S., Dienstmann, R., Margolin, A. A. and Guinney, J. (2014). Stepwise group sparse regression (SGSR): gene-set-based pharmacogenomic predictive models with stepwise selection of functional priors. In *Pacific Symposium on Biocomputing Co-Chairs* (pp. 32-43).

Jia, P., Kao, C.-F., Kuo, P.-H. and Zhao, Z. (2011). A comprehensive network and pathway analysis of candidate genes in major depressive disorder. *BMC Systems Biology, 5(Suppl 3),* p. S12.

Jia, P. and Zhao, Z. (2014). Network-assisted analysis to prioritize GWAS results: principles, methods and perspectives. *Human Genetics, 133(2),* pp. 125–138.

Johannes, M., Brase, J. C., Fröhlich, H., Gade, S., Gehrmann, M., Fälth, M., Sültmann, H. and Beißbarth, T. (2010). Integration of pathway knowledge into a reweighted recursive feature elimination approach for risk stratification of cancer patients. *Bioinformatics, 26(17),* pp. 2136–2144.

Jonsson, T. and Wheater, J. F. (1998). Area distribution for directed random walks. *Journal of statistical physics, 92(3–4),* pp. 713–725.

Kamat, P., Zhang, Y., Trappe, W. and Ozturk, C. (2005, June). Enhancing source-location privacy in sensor network routing. In *25th IEEE international conference on distributed computing systems (ICDCS'05)* (pp. 599-608). IEEE.

Kar, S., Sharma, K. D. and Maitra, M. (2016). A particle swarm optimization based gene identification technique for classification of cancer subgroups. In *2016 2nd International Conference on Control, Instrumentation, Energy and Communication (CIEC)* (pp. 130-134). IEEE.

Karaboga, D. and Akay, B. (2009). A survey: algorithms simulating bee swarm intelligence. *Artificial Intelligence Review, 31(1–4),* pp. 61–85.

Karn, T., Metzler, D., Ruckhäberle, E., Hanker, L., Gätje, R., Solbach, C., Ahr, A., Schmidt, M., Holtrich, U., Kaufmann, M. and Rody, A. (2010). Data driven derivation of cutoffs from a pool of 3,030 Affymetrix arrays to stratify distinct clinical types of breast cancer. *Breast Cancer Research and Treatment, 120(3),* pp. 567–579.

Karo, I. M. K., MaulanaAdhinugraha, K. and Huda, A. F. (2017, November). A cluster validity for spatial clustering based on Davies Bouldin index and Polygon Dissimilarity function. In *2017 Second International Conference on Informatics and Computing (ICIC)* (pp. 1-6). IEEE.

Kaufman, L. and Rousseeuw, P. (2009). *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons.

Kavzoglu, T., Sahin, E. K. and Colkesen, I. (2014). Landslide susceptibility mapping using GIS-based multi-criteria decision analysis, support vector machines, and logistic regression. *Landslides, 11(3),* pp. 425–439.

Khatri, P., Sirota, M. and Butte, A. J. (2012). Ten Years of Pathway Analysis: Current Approaches and Outstanding Challenges. *PLoS Computational Biology, 8(2),* e1002375.

Kim, E. K. and Choi, E. J. (2010). Pathological roles of MAPK signaling pathways in human diseases. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1802(4),* pp. 396–405.

Kim, S. Y. and Volsky, D. J. (2005). PAGE: parametric analysis of gene set enrichment. *BMC Bioinformatics, 6(1),* p. 144.

Kiselev, V. Y., Andrews, T. S. and Hemberg, M. (2019). Challenges in unsupervised clustering of single-cell RNA-seq data. *Nature Reviews Genetics, 20,* pp. 273–282.

Kittas, A., Delobelle, A., Schmitt, S., Breuhahn, K., Guziolowski, C. and Grabe, N. (2016). Directed random walks and constraint programming reveal active pathways in hepatocyte growth factor signaling. *FEBS Journal, 283(2),* pp. 350–360.

Knowles, J. D. and Corne, D. W. (2000). M-PAES: A memetic algorithm for multiobjective optimization. In *Proceedings of the 2000 Congress on Evolutionary Computation* (pp. 325–332). IEEE, Turkey.

Kothandan, R. and Biswas, S. (2015). Identifying microRNAs involved in cancer pathway using support vector machines. *Computational biology and chemistry, 55,* pp. 31–36.

Kourou, K., Papaloukas, C. and Fotiadis, D. I. (2016, February). Gene-based pathway enrichment analysis of oral squamous cell carcinoma patients. In *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI)* (pp. 360-363). IEEE.

Kriegel, H., Kröger, P., Sander, J. and Zimek, A. (2011). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 1(3),* pp. 231 – 240.

Kumar, P. and Wasan, S. K. (2011). Comparative study of k-means, pam and rough k-means algorithms using cancer datasets. In *Proceedings of CSIT: 2009 International Symposium on Computing, Communication, and Control (ISCCC 2009)* (Vol. 1, pp. 136-140). Singapore.

Kuner, R., Muley, T., Meister, M., Ruschhaupt, M., Buness, A., Xu, E. C., Schnabel, P., Warth, A., Poustka, A., Sültmann, H. and Hoffmann, H. (2009). Global gene expression analysis reveals specific patterns of cell junctions in non-small cell lung cancer subtypes. *Lung Cancer, 63(15),* pp. 32–38.

Labed, K., Fizazi, H., Mahi, H. and Galvan, I. M. (2018). A Comparative Study of Classical Clustering Method and Cuckoo Search Approach for Satellite Image Clustering: Application to Water Body Extraction. *Applied Artificial Intelligence, 32(1),* pp. 96–118.

Lai, C. Y., Tsai, P. F., Chang, S., Wang, Y. C. and Teng, L. W. (2017). The productivity opportunities by applying machine learning algorithms in a fab. In *2017 Joint International Symposium on e-Manufacturing and Design Collaboration (eMDC) and Semiconductor Manufacturing (ISSM)* (pp. 1-2). IEEE.

Li, Q., Yu, M. and Wang, S. (2017). A statistical framework for pathway and gene identification from integrative analysis. *Journal of multivariate analysis, 156,* pp. 1–17.

Li, T., Zhang, C. and Ogihara, M. (2004). A comparative study of feature selection and multiclass classification methods for tissue classification based on gene expression. *Bioinformatics, 20(15),* pp. 2429–2437.

Li, X., Peng, S., Zhan, X., Zhang, J. and Xu, Y. (2011). Comparison of feature selection methods for multiclass cancer classification based on microarray data. In *2011 4th International Conference on Biomedical Engineering and Informatics (BMEI)* (Vol. 3, pp. 1692-1696). IEEE.

Li, X., Shen, L., Shang, X. and Liu, W. (2015). Subpathway Analysis based on Signaling-Pathway Impact Analysis of Signaling Pathway. *PLoS ONE, 10(7),* e0132813.

Li, Y. and Patra, J. C. (2010). Integration of multiple data sources to prioritize candidate genes using discounted rating system. *BMC Bioinformatics, 11(SUPPLL.1)*.

Li, Y., Wang, G., Chen, H., Shi, L. and Qin, L. (2013). An Ant Colony Optimization Based Dimension Reduction Method for High-Dimensional Datasets. *Journal of Bionic Engineering, 10(2),* pp. 231–241.

Liu, C., Lehtonen, R. and Hautaniemi, S. (2018a). PerPAS: topology-based single sample pathway analysis method. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 15(3),* pp. 1022–1027.

Liu, D., Li, T. and Liang, D. (2014). Incorporating logistic regression to decision-theoretic rough sets for classifications. *International Journal of Approximate Reasoning, 55(1),* pp. 197–210.

Liu, H. C., Ma, F., Shen, Y., Hu, Y. Q. and Pan, S. (2015a). Overexpression of SMAR1 enhances radiosensitivity in human breast cancer cell line MCF7 via activation of p53 signaling pathway. *Oncology Research Featuring Preclinical and Clinical Cancer Therapeutics, 22(5–6),* pp. 293–300.

Liu, J., Lu, F., Gong, Y., Zhao, C., Pan, Q., Ballantyne, S., Zhao, X., Tian, S. and Chen, H. (2018b). High expression of synthesis of cytochrome c oxidase 2 and TP53-induced glycolysis and apoptosis regulator can predict poor prognosis in human lung. *Human pathology, 77,* pp. 54–62.

Liu, L., Wei, J. and Ruan, J. (2017a). Pathway enrichment analysis with networks. *Genes, 8(10),* p. 246.

Liu, W., Li, C., Xu, Y., Yang, H., Yao, Q., Han, J., Shang, D., Zhang, C., Su, F., Li, Xia, Li, Xiaoxi, Xiao, Y., Zhang, F. and Dai, M. (2013a). Topologically inferring risk-active pathways toward precise cancer classification by directed random walk. *Bioinformatics, 29(17),* pp. 2169–2177.

Liu, W., Wang, Q., Zhao, J., Zhang, C., Liu, Y., Zhang, J., Bai, X., Li, X., Feng, H., Liao, M. and Wang, W. (2015b). Integration of pathway structure information into a reweighted partial Cox regression approach for survival analysis on high-dimensional gene expression data. *Molecular bioSystems, 11(7),* pp. 1876–1886.

Liu, W., Wang, W., Tian, G., Xie, W., Lei, L., Liu, J., Huang, W., Xu, L. and Li, E. (2017b). Topologically inferring pathway activity for precise survival outcome prediction: breast cancer as a case. *Molecular bioSystems, 13(3),* pp. 537–548.

Liu, X. Y., Wu, J. and Zhou, Z. H. (2008). Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 39(2),* pp. 539–550.

Liu, Y., Li, Z., Xiong, H., Gao, X. and Wu, J. (2010, December). Understanding of internal clustering validation measures. In *2010 IEEE International Conference on Data Mining* (pp.

911-916). IEEE.

Liu, Y., Li, Z., Xiong, H., Gao, X., Wu, J. and Wu, S. (2013b). Understanding and enhancement of internal clustering validation measures. *IEEE transactions on cybernetics, 43(3),* pp. 982–994.

Liu, Z., Lu, Y., He, Z., Chen, L. and Lu, Y. (2015c). Expression analysis of the estrogen receptor target genes in renal cell carcinoma. *Molecular Medicine Reports, 11(1),* pp. 75–82.

Lix, L. M., Keselman, J. C. and Keselman, H. J. (1996). Consequences of Assumption Violations Revisited: A Quantitative Review of Alternatives to the One-Way Analysis of Variance F Test. *Review of Educational Research, 66(4),* pp. 579–619.

Lorena, A. C., De Carvalho, A. C. and Gama, J. M. (2008). A review on the combination of binary classifiers in multiclass problems. *Artificial Intelligence Review, 30(1–4),* p. 19.

Lovász, L. (1993). Random walks on graphs: A survey. *Paul erdos is eighty, 2(1),* pp. 1–46.

Lu, Y. and Han, J. (2003). Cancer classification using gene expression data. *Information Systems, 28(4),* pp. 243–268.

Ludwig, J. A. and Weinstein, J. N. (2005). Biomarkers in cancer staging, prognosis and treatment selection. *Nature Reviews Cancer, 5(11),* p. 845.

Ludwig, S. A., Picek, S. and Jakobovic, D. (2018). Classification of Cancer Data: Analyzing Gene Expression Data Using a Fuzzy Decision Tree Algorithm. In *Operations Research Applications in Health Care Management* (pp. 327-347). Springer, Cham.

Lynn, N., Ali, M. Z. and Suganthan, P. N. (2018). Population topologies for particle swarm optimization and differential evolution. *Swarm and Evolutionary Computation, 39*, pp. 24–35.

Ma, C., Chen, Y., Wilkins, D., Chen, X. and Zhang, J. (2015). An unsupervised learning approach to find ovarian cancer genes through integration of biological data. *BMC Genomics, 16(S9),* S3.

Macher, J. P. and Crocq, M. A. (2004). Treatment goals: response and nonresponse. *Dialogues in clinical neuroscience, 6(1),* p. 83.

Madhukar, N. S., Elemento, O. and Pandey, G. (2015). Prediction of Genetic Interactions Using Machine Learning and Network Properties. *Frontiers in Bioengineering and Biotechnology, 3,* p. 172.

Majhi, S. K. and Biswal, S. (2018). Optimal cluster analysis using hybrid K-Means and Ant Lion Optimizer. *Karbala International Journal of Modern Science, 4(4),* pp. 347–360.

Majhi, S. K. and Biswal, S. (2019). A Hybrid Clustering Algorithm Based on Kmeans and Ant Lion Optimization. In *Emerging Technologies in Data Mining and Information Security* (pp. 639-650). Springer, Singapore.

Makropoulou, M. (2016). *Cancer and electromagnetic radiation therapy: Quo Vadis?* Medical Physics.

Martini, P., Sales, G., Massa, M. S., Chiogna, M. and Romualdi, C. (2012). Along signal paths: an empirical gene set approach exploiting pathway topology. *Nucleic acids research, 41(1),* pp. e19–e19.

Mary, C. and Raja, S. K. (2009). Refinement of Clusters from K-

Means with Ant Colony Optimization. *Journal of Theoretical and Applied Information Technology, 10(1)*.

Mehmood, R., El-Ashram, S., Bie, R. and Sun, Y. (2018). Effective cancer subtyping by employing density peaks clustering by using gene expression microarray. *Personal and Ubiquitous Computing, 22(3)*, pp. 615–619.

Menashe, I., Figueroa, J. D., Garcia-Closas, M., Chatterjee, N., Malats, N., Picornell, A., Maeder, D., Yang, Q., Prokunina-Olsson, L., Wang, Z., Real, F. X., Jacobs, K. B., Baris, D., Thun, M., Albanes, D., Purdue, M. P., Kogevinas, M., Hutchinson, A., Fu, Y.-P., Tang, W., Burdette, L., Tardón, A., Serra, C., Carrato, A., García-Closas, R., Lloreta, J., Johnson, A., Schwenn, M., Schned, A., Andriole, G., Black, A., Jacobs, E. J., Diver, R. W., Gapstur, S. M., Weinstein, S. J., Virtamo, J., Caporaso, N. E., Landi, M. T., Fraumeni, J. F., Chanock, S. J., Silverman, D. T. and Rothman, N. (2012). Large-Scale Pathway-Based Analysis of Bladder Cancer Genome-Wide Association Data from Five Studies of European Background. *PLoS ONE, 7(1)*, e29396.

Mikaeil, R., Haghshenas, S. S. and Hoseinie, S. H. (2018). Rock penetrability classification using artificial bee colony (ABC) algorithm and self-organizing map. *Geotechnical and Geological Engineering, 36(2)*, pp. 1309–1318.

Mitra, A. P., Almal, A. A., George, B., Fry, D. W., Lenehan, P. F., Pagliarulo, V., Cote, R. J., Datar, R. H. and Worzel, W. P. (2006). The use of genetic programming in the analysis of quantitative gene expression profiles for identification of nodal status in bladder cancer. *BMC Cancer, 6(1)*, p. 159.

Mitrea, C., Taghavi, Z., Bokanizad, B., Hanoudi, S., Tagett, R., Donato, M., Voichiţa, C. and Drăghici, S. (2013). Methods and approaches in the topology-based analysis of biological pathways. *Frontiers in Physiology, 4,* p. 278.

Mohamad, M. S., Omatu, S., Deris, S. and Yoshioka, M. (2013a, April). A constraint and rule in an enhancement of binary particle swarm optimization to select informative genes for cancer classification. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining* (pp. 168-178). Springer, Berlin, Heidelberg.

Mohamad, M. S., Omatu, S., Deris, S., Yoshioka, M., Abdullah, A. and Ibrahim, Z. (2013b). An enhancement of binary particle swarm optimization for gene selection in classifying cancer classes. *Algorithms for Molecular Biology, 8(1),* p. 15.

Mohammadrezapour, O., Kisi, O. and Pourahmad, F. (2018). Fuzzy c-means and K-means clustering with genetic algorithm for identification of homogeneous regions of groundwater quality. *Neural Computing and Applications*. 1-13.

Mohammed, A., Biegert, G., Adamec, J. and Helikar, T. (2017). Identification of potential tissue-specific cancer biomarkers and development of cancer versus normal genomic classifiers. *Oncotarget, 8(49),* p. 85692.

Mohapatra, P., Chakravarty, S. and Dash, P. K. (2016). Microarray medical data classification using kernel ridge regression and modified cat swarm optimization based gene selection system. *Swarm and Evolutionary Computation, 28,* pp. 144–160.

Moorthy, K. and Mohamad, M. S. (2011, July). Random Forest for Gene Selection and Microarray Data Classification. In

Knowledge Technology Week (pp. 174-183). Springer, Berlin, Heidelberg.

Moreno-Torres, J. G., Sáez, J. A. and Herrera, F. (2012). Study on the Impact of Partition-Induced Dataset Shift on-Fold Cross-Validation. *IEEE Transactions on Neural Networks and Learning Systems, 23(8),* pp. 1304–1312.

Mortazavi, A., Toğan, V. and Moloodpoor, M. (2019). Solution of structural and mathematical optimization problems using a new hybrid swarm intelligence optimization algorithm. *Advances in Engineering Software, 127,* pp. 106–123.

Murohashi, M., Hinohara, K., Kuroda, M., Isagawa, T., Tsuji, S., Kobayashi, S., Umezawa, K., Tojo, A., Aburatani, H. and Gotoh, N. (2010). Gene set enrichment analysis provides insight into novel signalling pathways in breast cancer stem cells. *British journal of cancer, 102(1),* p. 206.

Murtagh, F. and Contreras, P. (2012). Algorithms for hierarchical clustering: an overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2(1),* pp. 86–97.

Mushtaq, H., Khawaja, S. G., Akram, M. U., Yasin, A., Muzammal, M., Khalid, S. and Khan, S. A. (2018). A Parallel Architecture for the Partitioning around Medoids (PAM) Algorithm for Scalable Multi-Core Processor Implementation with Applications in Healthcare. *Sensors, 18(12),* p. 4129.

Nacu, Ş., Critchley-Thorne, R., Lee, P. and Holmes, S. (2007). Gene expression network analysis and applications to immunology. *Bioinformatics, 23(7),* pp. 850–858.

Nagpal, A., Jatain, A. and Gaur, D. (2013, April). Review based on data clustering algorithms. In *2013 IEEE Conference on*

Information and Communication Technologies (pp. 298-303). IEEE.

Nair, R. P., Duffin, K. C., Helms, C., Ding, J., Stuart, P. E., Goldgar, D., Gudjonsson, J. E., Li, Y., Tejasvi, T., Feng, B. J. and Ruether, A. (2009). Genome-wide scan reveals association of psoriasis with IL-23 and NF-κB pathways. *Nature Genetics, 41(2),* p. 199.

Napier, N. and Limogiannis, N. (2016). A Bioinformatic Approach to MSI Cancer Research. *Bioengineering and Bioscience, 4(1),* pp. 7–10.

Nayak, J., Naik, B. and Behera, H. S. (2015). Fuzzy C-means (FCM) clustering algorithm: a decade review from 2000 to 2014. In *Computational intelligence in data mining-volume 2* (pp. 133-149). Springer, New Delhi.

Nazarenko, A. V. (2011). Directed random walk on the lattices of genus two. *International Journal of Modern Physics B, 25(26),* pp. 3415–3433.

Ng, R. T. and Han, J. (2002). CLARANS: A method for clustering objects for spatial data mining. *IEEE Transactions on Knowledge and Data Engineering, 56,* pp. 1003–1016.

Nies, H. W., Mohd Daud, K., Remli, M. A., Mohamad, M. S., Deris, S., Omatu, S., Kasim, S. and Sulong, G. (2017a, June). Classification of Colorectal Cancer Using Clustering and Feature Selection Approaches. In *International Conference on Practical Applications of Computational Biology and Bioinformatics* (pp. 58-65). Springer, Cham.

Nies, H. W., Zakaria, Z., Mohamad, M. S., Chan, W. H., Zaki, N., Sinnott, R. O., Napis, S., Chamoso, P., Omatu, S. and

Corchado, J. M. (2019). A Review of Computational Methods for Clustering Genes with Similar Biological Functions. *Processes, 7(9),* p. 550.

Nies, H. W. (2020). *Identification of Pathway and Gene Markers Using Enhanced Directed Random Walk for Multiclass Cancer Expression Data*. (Current Thesis)

Nies, Y. H., Islahudin, F., Chong, W. W., Abdullah, N., Ismail, F., Bustamam, R. S. A., Wong, Y. F., Saladina, J. J. and Shah, N. M. (2017b). Treatment decision-making among breast cancer patients in Malaysia. *Patient Preference and Adherence, 11,* pp. 1767–1777.

Nur, U., Shack, L. G., Rachet, B., Carpenter, J. R. and Coleman, M. P. (2009). Modelling relative survival in the presence of incomplete data: a tutorial. *International journal of epidemiology, 39(1),* pp. 118–128.

Obuchowski, N. A. and Bullen, J. A. (2018). Receiver operating characteristic (ROC) curves: review of methods with applications in diagnostic medicine. *Physics in Medicine and Biology, 63(7),* 07TR01.

Ortiz-Ramón, R., Larroza, A., Ruiz-España, S., Arana, E. and Moratal, D. (2018). Classifying brain metastases by their primary site of origin using a radiomics approach based on texture analysis: a feasibility study. *European Radiology, 28(11),* pp. 4514–4523.

Otukei, J. R. and Blaschke, T. (2010). Land cover change assessment using decision trees, support vector machines and maximum likelihood classification algorithms. *International Journal of Applied Earth Observation and Geoinformation, 12,* pp. S27–

S31.

Oyelade, J., Isewon, I., Oladipupo, F., Aromolaran, O., Uwoghiren, E., Ameh, F., Achas, M. and Adebiyi, E. (2016). Clustering Algorithms: Their Application to Gene Expression Data. *Bioinformatics and Biology Insights, 10,* BBI.S38316.

Pacheco, T. M., Gonçalves, L. B., Ströele, V. and Soares, S. S. R. (2018, July). An Ant Colony Optimization for Automatic Data Clustering Problem. In *2018 IEEE Congress on Evolutionary Computation (CEC)* (pp. 1-8). IEEE.

Patel, N., Jhadav, B., Aljouie, A. and Roshan, U. (2015, November). Cross-validation and cross-study validation of chronic lymphocytic leukemia with exome sequences and machine learning. In *2015 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1367-1374). IEEE.

Pawitan, Y., Bjöhle, J., Amler, L., Borg, A.-L., Egyhazi, S., Hall, P., Han, X., Holmberg, L., Huang, F., Klaar, S., c E. T., Miller, L., Nordgren, H., Ploner, A., Sandelin, K., Shaw, P. M., Smeds, J., Skoog, L., Wedrén, S. and Bergh, J. (2005). Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts. *Breast Cancer Research, 7(6),* p. R953.

Pécuchet, N., Popova, T., Manié, E., Lucchesi, C., Battistella, A., Vincent-Salomon, A., Caux-Moncoutier, V., Bollet, M., Sigal-Zafrani, B., Sastre-Garau, X., Stoppa-Lyonnet, D. and Stern, M.-H. (2013). Loss of heterozygosity at 13q13 and 14q32 predicts BRCA2 inactivation in luminal breast carcinomas. *International Journal of Cancer, 133(12),* 2834-

2842.

Peng, J., Zhu, L., Wang, Y. and Chen, J. (2019). Mining Relationships among Multiple Entities in Biological Networks. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB).*

Petrochilos, D., Shojaie, A., Gennari, J. and Abernethy, N. (2013). Using random walks to identify cancer-associated modules in expression data. *BioData Mining, 6(1),* p. 17.

Phongwattana, T., Engchuan, W. and Chan, J. H. (2015, January). Clustering-based multi-class classification of complex disease. In *2015 7th International Conference on Knowledge and Smart Technology (KST)* (pp. 25-29). IEEE.

Pilevar, A. H. and Sukumar, M. (2005). GCHL: A grid-clustering algorithm for high-dimensional very large spatial data bases. *Pattern Recognition Letters, 26(7),* pp. 999–1010.

Pillai, R. N., Behera, M., Berry, L. D., Rossi, M. R., Kris, M. G., Johnson, B. E., Bunn, P. A., Ramalingam, S. S. and Khuri, F. R. (2017). HER2 mutations in lung adenocarcinomas: A report from the Lung Cancer Mutation Consortium. *Cancer, 123(21),* pp. 4099–4105.

Quintela-Fandino, M., Arpaia, E., Brenner, D., Goh, T., Yeung, F. A., Blaser, H., Alexandrova, R., Lind, E. F., Tusche, M. W., Wakeham, A. and Ohashi, P. S. (2010). HUNK suppresses metastasis of basal type breast cancers by disrupting the interaction between PP2A and cofilin-1. *Proceedings of the National Academy of Sciences*, *107*(6), 2622-2627.

Rajkumar, P., Vennila, I. and Nirmalakumari, K. (2013). A novel hybrid method for gene selection in microarray based cancer classification. *International Journal of Engineering Science and Technology, 5(5),* p. 1104.

Rashedi, E., Nezamabadi-Pour, H. and Saryazdi, S. (2009). GSA: a gravitational search algorithm. *Information Sciences, 179(13),* pp. 2232–2248.

Re, M. and Valentini, G. (2012, September). Random Walking on Functional Interaction Networks to Rank Genes Involved in Cancer. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 66-75). Springer, Berlin, Heidelberg.

Rechkalov, T. V. (2015). Partition Around Medoids Clustering on the Intel Xeon Phi Many-Core Coprocessor. In *CEUR Workshop Proceedings. Vol. 1513: Proceedings of the 1st Ural Workshop on Parallel, Distributed, and Cloud Computing for Young Scientists (Ural-PDC 2015).—Yekaterinburg, 2015*.

Rejani, Y. I. A. and Selvi, S. T. (2009). Early Detection of Breast Cancer using SVM Classifier Technique. *International Journal on Computer Science and Engineering (IJCSE), 1(3),* pp. 127–130.

Remli, M. A., Mohd Daud, K., Nies, H. W., Mohamad, M. S., Deris, S., Omatu, S., Kasim, S. and Sulong, G. (2017, June). K-Means Clustering with Infinite Feature Selection for Classification Tasks in Gene Expression Data. In *International Conference on Practical Applications of Computational Biology and Bioinformatics* (pp. 50-57). Springer, Cham.

Roberts, M. and Russo, R. (2014). *A student's guide to analysis of variance*. 1st Edition. London: Routledge.

Rodriguez, J. D., Perez, A. and Lozano, J. A. (2009). Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence, 32(3),* pp. 569–575.

Ross, A. and Willson, V. L. (2017). One-Way ANOVA. *Basic and Advanced Statistical Tests*, pp. 21–24.

Roux, M. (2018). A Comparative Study of Divisive and Agglomerative Hierarchical Clustering Algorithms. *Journal of Classification, 35(2),* pp. 345–366.

Roy, S., Shah, V. K. and Das, S. K. (2019). Design of Robust and Efficient Topology using Enhanced Gene Regulatory Networks. *IEEE Transactions on Molecular, Biological and Multi-Scale Communications.*

Sáez, A., Sánchez-Monedero, J., Gutiérrez, P. A. and Hervás-Martínez, C. (2015). Machine learning methods for binary and multiclass classification of melanoma thickness from dermoscopic images. *IEEE transactions on medical imaging, 35(4),* pp. 1036–1045.

Sahoo, G. and Kumar, Y. (2012). Analysis of parametric and non parametric classifiers for classification technique using WEKA. *International Journal of Information Technology and Computer Science (IJITCS), 4(7),* p. 43.

Sahu, V., Mohan, A. and Dey, S. (2019). p38 MAP kinases: plausible diagnostic and prognostic serum protein marker of non small cell lung cancer. *Experimental and Molecular Pathology. Academic Press, 107,* pp. 118–123.

Sanchez-Palencia, A., Gomez-Morales, M., Gomez-Capilla, J. A., Pedraza, V., Boyero, L., Rosell, R. and Fárez-Vidal, M. E.

(2011). Gene expression profiling reveals novel biomarkers in nonsmall cell lung cancer. *International Journal of Cancer, 129(2),* pp. 355–364.

Santhisree, K. and Damodaram, A. (2011, April). CLIQUE: Clustering based on density on web usage data: Experiments and test results. In *2011 3rd International Conference on Electronics Computer Technology* (Vol. 4, pp. 233-236). IEEE.

Schaffer, C. (1993). Selecting a classification method by cross-validation. *Machine Learning, 13(1),* pp. 135–143.

Schoenborn, N. L., Xue, Q. L., Pollack, C. E., Janssen, E. M., Bridges, J. F., Wolff, A. C. and Boyd, C. M. (2019). Demographic, health, and attitudinal factors predictive of cancer screening decisions in older adults. *Preventive medicine reports, 13,* pp. 244–248.

Seah, C. S., Kasim, S., Fudzee, M. F. M., Ping, J. M. L. T., Mohamad, M. S., Saedudin, R. R. and Ismail, M. A. (2017). An enhanced topologically significant directed random walk in cancer classification using gene expression datasets. *Saudi journal of biological sciences, 24(8),* pp. 1828–1841.

Shanmugam, C. and Sekaran, E. C. (2019). IRT image segmentation and enhancement using FCM-MALO approach. *Infrared Physics and Technology,* pp. 187–196.

Shchur, L. N., Heringa, J. R. and Blöte, H. W. J. (1997). Simulation of a directed random-walk model The effect of pseudo-random-number correlations. *Physica A: Statistical Mechanics and its Applications, 241(3–4),* pp. 579–592.

Shen, L. and Tan, E. C. (2005). Dimension reduction-based penalized logistic regression for cancer classification using microarray data. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 2(2)*, pp. 166–175.

Shi, K., Gao, L. and Wang, B. (2018, June). Inferring Dysregulated Pathways of Driving Cancer Subtypes Through Multi-omics Integration. In *International Symposium on Bioinformatics Research and Applications* (pp. 101-112). Springer, Cham.

Sootanan, P., Meechai, A., Prom-on, S. and Chan, J. H. (2011, November). Pathway-Based Microarray Analysis with Negatively Correlated Feature Sets for Disease Classification. In *International Conference on Neural Information Processing* (pp. 676-683). Springer, Berlin, Heidelberg.

Srivastava, A., Chakrabarti, S., Das, S., Ghosh, S. and Jayaraman, V. K. (2013). Hybrid Firefly Based Simultaneous Gene Selection and Cancer Classification Using Support Vector Machines and Random Forests. In *Proceedings of seventh international conference on bio-inspired computing: theories and applications (BIC-TA 2012)* (pp. 485-494). Springer, India.

Su, J., Yoon, B.-J. and Dougherty, E. R. (2010). Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network. *BMC Bioinformatics, 11(6),* p. S8.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R., Lander, E. S. and Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences, 102(43),* pp. 15545–15550.

Sugiyama, M., Yamada, M., Kimura, M. and Hachiya, H. (2011). On Information-Maximization Clustering: Tuning Parameter Selection and Analytic Solution. In *Proceedings of the 28th International Conference on Machine Learning* (pp. 65-72). USA.

Sun, G., Shan, M. H., Ma, B. L., Geng, Z. L., Alibiyati, A., Zhong, H., Wang, J., Ren, G. H., Li, H. T. and Dong, C. (2012). Identifying crosstalk of mTOR signaling pathway of lobular breast carcinomas. *European Review for Medical and Pharmacological Sciences, 16(10),* pp. 1355–1361.

Swindell, W. R., Johnston, A., Carbajal, S., Han, G., Wohn, C., Lu, J., Xing, X., Nair, R. P., Voorhees, J. J., Elder, J. T., Wang, X.-J., Sano, S., Prens, E. P., DiGiovanni, J., Pittelkow, M. R., Ward, N. L. and Gudjonsson, J. E. (2011). Genome-Wide Expression Profiling of Five Mouse Models Identifies Similarities and Differences with Human Psoriasis. *PLoS ONE, 6(4),* e18266.

Tamposis, I. A., Tsirigos, K. D., Theodoropoulou, M. C., Kontou, P. I., Tsaousis, G. N., Sarantopoulou, D., Litou, Z. I. and Bagos, P. G. (2019). JUCHMME: A Java Utility for Class Hidden Markov Models and Extensions for biological sequence analysis. *Bioinformatics,* pp. 1–4.

Tang, H., Zeng, T. and Chen, L. (2019). High-Order Correlation Integration for Single-Cell or Bulk RNA-seq Data Analysis. *Frontiers in Genetics,* 10.

Tarca, A. L., Lauria, M., Unger, M., Bilal, E., Boue, S., Kumar Dey, K., Hoeng, J., Koeppl, H., Martin, F., Meyer, P. and Nandy, P. (2013). Strengths and limitations of microarray-based phenotype prediction: lessons learned from the IMPROVER Diagnostic Signature Challenge. *Bioinformatics, 29(22),* pp. 2892–2899.

Tharwat, A. and Hassanien, A. E. (2019). Quantum-Behaved Particle Swarm Optimization for Parameter Optimization of Support Vector Machine. *Journal of Classification,* pp. 1–23.

Tian, J. and Gu, M. (2019). Subspace Clustering Based on Self-organizing Map. In *Proceeding of the 24th International Conference on Industrial Engineering and Engineering Management 2018* (pp. 151-159). Springer, Singapore.

Tian, S., Chang, H. H. and Wang, C. (2016). Weighted-SAMGSR: combining significance analysis of microarray-gene set reduction algorithm with pathway topology-based weights to select relevant genes. *Biology Direct, 11(1),* p. 50.

Usoskin, D., Furlan, A., Islam, S., Abdo, H., Lönnerberg, P., Lou, D., Hjerling-Leffler, J., Haeggström, J., Kharchenko, O., Kharchenko, P. V. and Linnarsson, S. (2015). Unbiased classification of sensory neuron types by large-scale single-cell RNA sequencing. *Nature neuroscience, 18(1),* p. 145.

Van De Leemput, J., Boles, N. C., Kiehl, T. R., Corneo, B., Lederman, P., Menon, V., Lee, C., Martinez, R. A., Levi, B. P., Thompson, C. L., Yao, S., Kaykas, A., Temple, S. and Fasano, C. A. (2014). CORTECON: A Temporal Transcriptome Analysis of In Vitro Human Cerebral Cortex Development from Human Embryonic Stem Cells. *Neuron, 83(1),* pp. 51–68.

Vaske, C. J., Benz, S. C., Sanborn, J. Z., Earl, D., Szeto, C., Zhu, J., Haussler, D. and Stuart, J. M. (2010). Inference of patientspecific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics. 26(12),* i237–i245.

Vijendra, S. (2011). Efficient clustering for high dimensional data: Subspace based clustering and density based clustering. *Information Technology Journal, 10(6),* pp. 1092–1105.

Vrana, K. E., Freeman, W. M. and Aschner, M. (2003). Use of microarray technologies in toxicology research. *Neurotoxicology, 24(3),* pp. 321–332.

Wang, B., Zhang, L. and Gong, N. Z. (2017, May). SybilSCAR: Sybil detection in online social networks via local rule based propagation. In *IEEE INFOCOM 2017-IEEE Conference on Computer Communications* (pp. 1-9). IEEE.

Wang, J., Zhu, C., Zhou, Y., Zhu, X., Wang, Y. and Zhang, W. (2018a). From partition-based clustering to density-based clustering: Fast find clusters with diverse shapes and densities in spatial databases. *IEEE Access, 6,* pp. 1718–1729.

Wang, J., Zuo, Y., Man, Y. G., Avital, I., Stojadinovic, A., Liu, M., Yang, X., Varghese, R. S., Tadesse, M. G. and Ressom, H. W. (2015). Pathway and network approaches for identification of cancer signature markers from omics data. *Journal of Cancer, 6(1),* p. 54.

Wang, K., Li, M. and Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics, 11(12),* p. 843.

Wang, R., Cai, Y., Zhang, B. and Wu, Z. (2018b). A 16-gene expression signature to distinguish stage I from stage II lung squamous carcinoma. *International journal of molecular* *medicine, 41(3),* pp. 1377–1384.

Wang, W. and Liu, W. (2018). Integration of gene interaction information into a reweighted random survival forest approach for accurate survival prediction and survival biomarker discovery. *Scientific reports, 8(1),* p. 13202.

Wang, W., Yang, J. and Muntz, R. (1997, August). STING: A statistical information grid approach to spatial data mining. In *Proceedings of 23rd International Conference on Very Large Data Bases (VLDB)* (Vol. 97, pp. 186-195). Athens, Greece.

Wang, Y., Tetko, I. V., Hall, M. A., Frank, E., Facius, A., Mayer, K. F. and Mewes, H. W. (2005). Gene selection from microarray data for cancer classification-a machine learning approach. *Computational biology and chemistry, 29(1),* pp. 37–46.

Wei, H. and Zheng, H. R. (2015). A signaling pathway analysis method based on information divergence. In *12th International Symposium on Operations Research and its Applications in Engineering, Technology and Management (ISORA 2015).*

Wei, Z. and Li, H. (2007). A Markov random field model for network-based analysis of genomic data. *Bioinformatics, 23(12),* pp. 1537–1544.

Xu, D. and Tian, Y. (2015). A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science, 2(2),* pp. 165–193.

Xu, R. and Wunsch, D. C. (2010). Clustering algorithms in biomedical research: a review. *IEEE reviews in biomedical engineering, 3,* pp. 120–154.

Xu, X., Li, J., Zhou, M., Xu, J. and Cao, J. (2018). Accelerated two-stage particle swarm optimization for clustering not-wellseparated data. *IEEE Transactions on Systems, Man, and Cybernetics: Systems, 99,* pp. 1–12.

Yang, L., Ainali, C., Tsoka, S. and Papageorgiou, L. G. (2014). Pathway activity inference for multiclass disease classification through a mathematical programming optimisation framework. *BMC Bioinformatics, 15(1),* p. 390.

Yang, Q., Wang, S., Dai, E., Zhou, S., Liu, D., Liu, H., Meng, Q., Jiang, B. and Jiang, W. (2017). Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. *Briefings in bioinformatics, 20(1),* pp. 168–177.

Yang, R., Daigle, B. J., Petzold, L. R. and Doyle, F. J. (2012). Core module biomarker identification with network exploration for breast cancer metastasis. *BMC Bioinformatics, 13(1),* p. 12.

Yang, S. and Naiman, D. Q. (2014). Multiclass cancer classification based on gene expression comparison. *Statistical applications in genetics and molecular biology*, *13*(4), 477-496.

Yao, Y., Richman, L., Morehouse, C., de los Reyes, M., Higgs, B. W., Boutrin, A., White, B., Coyle, A., Krueger, J., Kiener, P. A. and Jallal, B. (2008). Type I Interferon: Potential Therapeutic Target for Psoriasis?. *PLoS ONE, 3(7),* e2737.

Yasrebi, H., Sperisen, P., Praz, V. and Bucher, P. (2009). Can Survival Prediction Be Improved By Merging Gene Expression Data Sets?. *PLoS ONE, 4(10),* e7431.

Yazdani, S., Nezamabadi-pour, H. and Kamyab, S. (2014). A gravitational search algorithm for multimodal optimization. *Swarm and Evolutionary Computation, 14,* pp. 1–4.

Ye, S., Huang, X., Teng, Y. and Li, Y. (2018, March). K-means clustering algorithm based on improved Cuckoo search algorithm and its application. In *2018 IEEE 3rd international conference on big data analysis (ICBDA)* (pp. 422-426). IEEE.

Yeung, K. Y. and Bumgarner, R. E. (2003). Multiclass classification of microarray data with repeated measurements: application to cancer. *Genome biology, 4(12),* R83.

Yu, X., Yu, G. and Wang, J. (2017). Clustering cancer gene expression data by projective clustering ensemble. *PLOS ONE, 12(2),* e0171429.

Zelnik-Manor, L. and Perona, P. (2005). Self-tuning spectral clustering. *Advances in neural information processing systems,* pp. 1601–1608.

Zhang, H., Raitoharju, J., Kiranyaz, S. and Gabbouj, M. (2016). Limited random walk algorithm for big graph data clustering. *Journal of Big Data, 3(1),* p. 26.

Zhang, L. (2006, July). A self-adjusting directed random walk approach for enhancing source-location privacy in sensor network routing. In *Proceedings of the 2006 international conference on Wireless communications and mobile computing* (pp. 33-38).

Zhao, L., Lee, V. H., Ng, M. K., Yan, H. and Bijlsma, M. F. (2018). Molecular subtyping of cancer: current status and moving toward clinical applications. *Briefings in Bioinformatics, 20(2),* pp. 572–584.

Zhao, X., Sala, A., Zheng, H. and Zhao, B. Y. (2011, October). Efficient shortest paths on massive social graphs. In *7th International Conference on Collaborative Computing:*

Networking, Applications and Worksharing (CollaborateCom) (pp. 77-86). IEEE.

Zhe, S., Naqvi, S. A., Yang, Y. and Qi, Y. (2013). Joint network and node selection for pathway-based genomic data analysis. *Bioinformatics, 29(16),* pp. 1987–1996.

Zhou, J. and Fu, B. (2018). The research on gene-disease association based on text-mining of PubMed. *BMC Bioinformatics, 19(1),* p. 37.

Zhu, H. and Li, L. (2011). Biological pathway selection through nonlinear dimension reduction. *Biostatistics, 12(3),* pp. 429–444.

Zhu, L., Su, F., Xu, Y. and Zou, Q. (2018). Network-based method for mining novel HPV infection related genes using random walk with restart algorithm. *Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1864(6),* pp. 2376–2383.

Zou, R., Zhang, D., Lv, L., Shi, W., Song, Z., Yi, B., Lai, B., Chen, Q., Yang, S. and Hua, P. (2019). Bioinformatic gene analysis for potential biomarkers and therapeutic targets of atrial fibrillation-related stroke. *Journal of Translational Medicine, 17(1),* p. 45.

Zuo, Y., Cui, Y., Yu, G., Li, R. and Ressom, H. W. (2017). Incorporating prior biological knowledge for network-based differential gene expression analysis using differentially weighted graphical LASSO. *BMC Bioinformatics, 18(1),* p. 99.

LIST OF PUBLICATIONS

Journal with Impact Factor

1. **Nies, H. W.,** Zakaria, Z., Mohamad, M. S., Chan, W. H., Zaki, N., Sinnott, R. O., Napis, S., Chamoso, P., Omatu, S. and Corchado, J. M. (2019). A Review of Computational Methods for Clustering Genes with Similar Biological Functions. *Processes*, *7*(9), 550. **(Q2, IF:1.963)**
2. **Nies, H. W.**, Zakaria, Z., Mohamad, M. S., Chan, W. H., Zaki, N., and Ibrahim, Z. **(under review).** An Enhanced Directed Random Walk Method to Identify Pathway and Gene Markers for Multiclass Breast Cancer Expression Data. *Computer Methods and Programs in Biomedicine*. **(Q1, IF: 3.424)**

Indexed Conference Proceedings

1. **Nies, H. W.**, Mohd Daud, K., Remli, M. A., Mohamad, M. S., Deris, S., Omatu, S., Kasim, S. and Sulong, G. (2017). Classification of Colorectal Cancer Using Clustering and Feature Selection Approaches. In *International Conference on Practical Applications of Computational Biology and Bioinformatics* (pp. 58-65). Springer, Cham. **(Indexed by SCOPUS)**
2. Remli, M. A., Daud, K. M., **Nies, H. W.**, Mohamad, M. S., Deris, S., Omatu, S., Kasim, S. and Sulong, G. (2017, June). K-means clustering with infinite feature selection for classification tasks in gene expression data. In *International Conference on Practical Applications of Computational Biology and Bioinformatic* (pp. 50-57). Springer, Cham. **(Indexed by SCOPUS)**

Non-Indexed Conference Proceedings

1. **Nies, H.W.**, Zakaria, Z., Mohamad, M.S., A. Samah, A., Chan, W. H., and Deris, S. (2018). Review on Pathway Topology-Based Microarray Analysis. In *7th International Graduate Conference on Engineering Science and Humanity 2018*.

3. **Nies, H.W.**, Zakaria, Z., Mohamad, M.S., A. Samah, A., Chan, W. H., and Deris, S., (2018). Review on Weighting Category of Pathway Topology-Based Microarray Analysis in Multiclass Classification. *UTM Computing Proceedings.*

Copyrights

1. **Nies, H. W.**, Zakaria, Z., Chan, W. H (2020). *Automation of PubMed Text Data Mining for Biological Validation of Genes and Pathways*. Universiti Teknologi Malaysia.

2. **Nies, H. W.**, Zakaria, Z., Chan, W. H (2020). *Integration of K-Means Clustering and Average Silhouette Method into Directed Random Walk to Increase the Efficiency of Identifying Informative Genes in the Directed Graph*. Universiti Teknologi Malaysia.

3. **Nies, H. W.**, Zakaria, Z., Chan, W. H (2020). *An Enhanced Directed Random Walk Method to Identify Informative Genes and Pathways from Multiclass Cancer Expression Data*. Universiti Teknologi Malaysia.