IDENTIFICATION OF POTENTIAL BIOMARKERS FOR ESOPHAGEAL CANCER FROM GENE EXPRESSION AND INTERACTIONS USING BICLUSTERING ALGORITHM

GUI YU XUAN

UNIVERSITI TEKNOLOGI MALAYSIA

PSZ 19:16 (Pind. 1/13)

UNIVERSITI TEKNOLOGI MALAYSIA

DECLARATION OF THESIS / UNDERGRADUATE PROJECT REPORT AND COPYRIGHT

Author's full name : GUI YU XUAN

Date of Birth : 07 – 04 – 2000

Title : IDENTIFICATION OF POTENTIAL BIOMARKERS FOR ESOPHAGEAL CANCER FROM GENE EXPRESSION AND INTERACTIONS USING BICLUSTERING ALGORITHM

Academic Session : 20222023

I declare that this thesis is classified as:

☐	**CONFIDENTIAL**	(Contains confidential information under the Official Secret Act 1972)*
☐	**RESTRICTED**	(Contains restricted information as specified by the organization where research was done)*
☑	**OPEN ACCESS**	I agree that my thesis to be published as online open access (full text)

1. I acknowledged that Universiti Teknologi Malaysia reserves the right as follows:
2. The thesis is the property of Universiti Teknologi Malaysia
3. The Library of Universiti Teknologi Malaysia has the right to make copies for the purpose of research only.
4. The Library has the right to make copies of the thesis for academic exchange.

Certified by:

SIGNATURE OF STUDENT	**SIGNATURE OF SUPERVISOR**
A20EC0039	DR. CHAN WENG HOWE
MATRIC NUMBER	**NAME OF SUPERVISOR**
Date: 20 JUNE 2023	Date: 20 JUNE 2023

NOTES : If the thesis is CONFIDENTIAL or RESTRICTED, please attach with the letter from the organization with period and reasons for confidentiality or restriction

"I hereby declare that I have read this thesis and in my opinion this thesis is sufficient in term of scope and quality for the award of the degree of Bachelor of Computing Science (Bioinformatics)."

Signature	:	_____
Name of Supervisor I	:	DR CHAN WENG HOWE
Date	:	20 JUNE 2023

IDENTIFICATION OF POTENTIAL BIOMARKERS FOR ESOPHAGEAL CANCER FROM GENE EXPRESSION AND INTERACTIONS USING BICLUSTERING ALGORITHM

GUI YU XUAN

A thesis submitted in partial fulfilment of the
requirements for the award of the degree of
Bachelor of Computing Science (Bioinformatics)

Faculty of Computing
Universiti Teknologi Malaysia

JUNE 2023

DECLARATION

I declare that this thesis entitled *"Identification of Potential Biomarkers for Esophageal Cancer from Gene Expression and Interactions Using Biclustering Algorithm"* is the result of my own research except as cited in the references. The thesis has not been accepted for any degree and is not concurrently submitted in candidature of any other degree.

Signature : .....

Name : GUI YU XUAN

Date : 20 JUNE 2023

DEDICATION

This thesis is dedicated to my father, who taught me that the best kind of knowledge to have is that which is learned for its own sake. It is also dedicated to my mother, who taught me that even the largest task can be accomplished if it is done one step at a time.

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, lecturers and friend. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, Professor Dr. Chan Weng Howe, for encouragement, guidance, critics and friendship. Without his continued support and interest, this thesis would not have been the same as presented here.

I am also indebted to Universiti Teknologi Malaysia (UTM) for funding my Bachelor study. Librarians at UTM also deserve special thanks for their assistance in supplying the relevant literatures.

My fellow undergraduate student should also be recognised for their support. My sincere appreciation also extends to all my course mate and others who have provided assistance at various occasions. Their views and tips are useful indeed. Unfortunately, it is not possible to list all of them in this limited space. I am grateful to all my family member.

ABSTRACT

Biclustering is a strong data mining approach to group the clusters based on specific characteristic. There are different biclustering methods had been proposed to identify the potential biomarkers for a certain disease. However, most of the research are done based on the statistical data which may produce false positive result and overfit the data. Therefore, the lack of biological relevance data in biclustering analysis leads to low precision in identifying relevant gene clusters and decreases the accuracy of biomarkers detection. The purpose of this study is to propose a biclustering method to identify the potential biomarkers of esophageal cancer from gene expression data and PPI. In this research, the gene expression dataset and protein-protein interaction datasets will be undergo the gene selection process and use for biclustering method. Elbow method had been used to determine the optimum number of biclusters. After a number of biclusters had been obtained, each bicluster will then be classified by classification model which is Support Vector Machine. From the result obtained, the genes will then be verified by the biological knowledgebases for the confirmation. To be illustrate, the biclustering method that used in this research is Plaid model, which will select the rows and columns randomly from the dataset to form biclusters. A collection of biclusters will then be categorized to different group for further classification purpose. The group of bicluster with higher accuracy will then be validate with biological knowledgebases. The potential biomarkers for the esophageal cancer should include ARPC2, APPL1, FTL, and PLAU. To conclude, since the potential biomarkers for esophageal cancer can be found in this research, it has the potential to improve the early detection and diagnosis for the esophageal cancer and improve in the available treatments.

ABSTRAK

Biclustering ialah pendekatan perlombongan data yang berkuasa kepada kumpulan kumpulan berdasarkan ciri khusus. Terdapat beberapa kaedah biclustering yang telah dicadangkan untuk mengenal pasti biomarker yang berpotensi untuk penyakit tertentu. Walau bagaimanapun, kebanyakan kajian dijalankan menggunakan data statistik yang mungkin menghasilkan keputusan yang positif palsu dan terlalu menekankan data. Oleh itu, kekurangan data biologi dalam analisis biclustering menyebabkan ketepatan yang rendah dalam mengenal pasti kelompok gen yang berkaitan dan mengurangkan ketepatan pengesanan biomarker. Tujuan kajian ini adalah untuk mencadangkan kaedah biclustering dalam mengenal pasti biomarker berpotensi untuk kanser esofagus daripada data ekspresi gen dan interaksi protein-protein. Dalam penyelidikan ini, set data ekspresi gen dan set data interaksi protein-protein akan dijalankan pemilihan gen dan digunakan dalam kaedah biclustering. Kaedah Elbow digunakan untuk menentukan bilangan bicluster yang optimum. Setelah bicluster diperolehi, setiap bicluster akan dikelaskan menggunakan model klasifikasi yang bername Support Vector Machine. Berdasarkan keputusan yang diperoleh, gen kemudiannya akan disahkan menggunakan pangkalan pengetahuan biologi untuk pengesahan. Sebagai contoh, kaedah biclustering yang digunakan dalam kajian ini ialah model Plaid, yang memilih secara rawak baris dan lajur daripada input untuk membentuk bicluster. Koleksi bicluster kemudiannya akan dikategorikan kepada kumpulan yang berbeza untuk tujuan pengelasan selanjutnya. Bicluster dengan ketepatan yang lebih tinggi kemudiannya akan disahkan menggunakan pangkalan pengetahuan biologi. Biomarker berpotensi untuk kanser esofagus yang telah ditemui termasuk ARPC2, APPL1, FTL dan PLAU. Untuk membuat kesimpulan, penyelidikan ini menunjukkan bahawa biomarker berpotensi untuk kanser esofagus boleh ditemui, yang berpotensi untuk meningkatkan pengesanan awal dan diagnosis kanser esofagus dan menambah baik rawatan yang tersedia.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

PPI	-	Protein – Protein Interactions
EC	-	Esophageal Cancer
ESCC	-	Esophageal Squamous Cell Carcinoma
EAC	-	Esophageal Adenocarcinoma
OPSM	-	Order Preserving Submatrix
CPB	-	Correlated Pattern Biclustering
BBC	-	Bayesian Biclustering
ISA	-	Iterative Signature Algorithm
CC	-	Cheng & Church
MCMC	-	Markov chain Monte Carlo
BiMax	-	Binary inclusion-Maximal
GEO	-	Gene Expression Omnibus
STRING	-	Search Tool for the Retrieval of Interacting Genes/Proteins
SVM	-	Support Vector Machine
kNN	-	K-Nearest Neighbours
ANN	-	Artificial Neural Network
NCBI	-	National Centre for Biotechnology Information
UniProt	-	Universal Protein Resource
TP	-	True Positive
FP	-	False Positive
TN	-	True Negative
FN	-	False Negative
SSE	-	Sum of Square Error

LIST OF SYMBOLS

Z	-	Standard score
x	-	Observed value
μ	-	Mean of the samples
δ	-	Standard deviation of the samples
∑	-	Summation
X_i	-	Mean value of *i*th data
$\bar{X}$	-	Mean value for all data
	-	

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Introduction

Esophageal cancer (EC) is the world's eighth most frequent cancer (World Cancer Research Fund International, no date). EC is a type of cancer that develops in esophagus. Due to a lack of early symptoms, the diagnosis occurs in the middle and late stages and the risk of recurrence after therapy is significant causing the 5-year survival rate for EC is still poor (Wan, Smith and Wei, 2018). According to Karimizadeh et al (2019), the identification of molecular pathways and complicated disease mechanisms can be facilitated by combining different biological data useful to certain biological queries, which can also boost the accuracy of results. By performing gene expression analysis, thousands of genes' levels of expression in a tissue or cell type are simultaneously measured (Karimizadeh et al, 2019). Gene expression data give information about the levels of gene activity but do not fully capture the complexity of biological systems (Karimizadeh et al, 2019). By focusing just on gene expression, we run the risk of ignoring significant regulatory processes and missing important information required for a complete understanding. Hence, in order to have a full understanding on the connection between genes' activity, several data had been applying together with gene expression such as genomic data, proteomic data, metabolomics data and protein-protein interaction (PPI). Applying conserved pathways and protein complexes, alignment and mapping of PPI networks offers a chance to learn more about the evolutionary links across species (Athanasios, 2017). Additionally, it has been demonstrated that within sequence homology clusters, information from protein-protein interaction networks can predict functional orthologous proteins (Athanasios, 2017). As a result, the integration of information on PPI and gene expression enables the discovery of possible biomarkers and advances our understanding of disease.

According to National Cancer Institute, a biomarker is a biological molecule that can be detected in tissues, body fluids, or blood that can indicate if a certain process, condition, or disease is normal or pathological (National Cancer Institute, no date). The body's reaction to a sickness or condition's therapy can be monitored using biomarkers. Hence, by identifying biomarkers for EC have the potential to lower morbidity and death. Machine learning methods are a viable alternative to traditional data analysis approaches and be widely used in the biomarker discovery since they automatically discover patterns and relationships from data without explicit programming (Xie et al, 2021). Supervised learning such as decision trees, naïve bayes and neural network, unsupervised learning such as K-means clustering are the methods that available in the machine learning. For your information, biclustering is a strong data mining approach that enables grouping of rows and columns concurrently in a matrix format dataset (Xie et al, 2019). Biclustering methods are useful for analysing gene expression and PPI data because they identify sections of genes with comparable expression patterns across sample subsets or situations (Xie et al, 2019). Identifying the subsets of genes by combining genes and samples based on their expression patterns able to reduce the complexity of large datasets and identify networks of related genes that are co-expressed in specific sample subsets. Therefore, the biclustering method is a useful tool that can be used to analyse esophageal cancers through the gene expression data and PPI data to detect gene clusters that exhibit differential expression when compared to normal tissue in esophageal tumours.

1.2 Problem Background

The pattern of gene expression in a cell or tissue dictates its form and function. While there's over a thousand genes on a microarray chip, there are only a few samples. As a result, the curse of dimensionality, noise, and randomness of this data are significant issues that arise in the interpretation of microarray data and present numerous data mining and machine learning obstacles (Moteghae, Maghooli and Garshasbi, 2018). However, biclustering can decrease the high-dimensional character of gene expression datasets by focusing on these co-expressed genes, which can increase classification accuracy by decreasing noise and highlighting pertinent

features. In order to find gene networks that are co-expressed in a certain subset of samples, biclustering can be used to simultaneously group genes and samples based on their patterns of expression.

Even biclustering algorithms can be a useful tool to identify the relevant genes to improve the classification's accuracy, but biclustering algorithms have limitations in gene expression datasets. According to Eren et al. (2013), synthetic datasets frequently don't perform as well as gene expression datasets. At the same time, the performance of each algorithm varies depending on the circumstances bicluster model. Hence, it is necessary to consider the data and parameters used before choosing a biclustering algorithm.

1.3 Problem Statement

EC is extremely aggressive (Napier, Scheerer and Misra, 2014). Early detection of esophageal cancer able to produce effective patient outcome despite improvements in available treatments (Rai, Abdo and Agrawal, 2023). PPI and gene expression data can be used to identify potential EC biomarkers, which could help with early detection and the creation of targeted treatments. However, using only statistical significance data to find biomarkers can produce false-positive results and overfit the data (Rashidi et al, 2022). This is due to statistical significance data may obtaining an observation in which there is no relationship among the variables. As a result, we must determine the biological significance of the data to increase the possibility of discovering a true and informative biomarker. PPI and gene expression data are biological relevance data because they provide the interactions between genes and show the pattern of gene expression (Rao et al, 2014; National Human Genome Research Institute, 2023). To increase the precision of biomarker detection, biclustering algorithms have offer a solution to identify the co-expressed genes (Branders, Schaus and Dupont, 2019). Therefore, the problem statement of this study is that the lack of the biological relevance data in biclustering analysis leading to low precision in identifying relevant gene clusters and decrease the accuracy of biomarkers detection.

1.4 Research Goal

The goal of this research is to propose a biclustering method to identify the potential biomarkers of esophageal cancer from gene expression data and PPI.

1.5 Research Objectives

The objectives of the research are:

(a) To generate input data from gene expression, protein-protein interaction and pathway data.

(b) To implement biclustering algorithm in identification of potential biomarkers from the input data of gene expression.

(c) To evaluate the selected potential biomarkers using Support Vector Machine by creating the features to the input data of gene expression and splitting them into train and test data.

(d) To verify the identified potential biomarkers with biological knowledgebases such as NCBI and UniProt by checking if the identified biomarkers are known to be involved in the disease.

1.6 Research Scope

The scopes of the research are:

(a) Concentrate on a plaid biclustering method to identify esophageal cancer biomarkers.

(b) Programming languages for the study are Python and R.

(c) Esophageal cancer data retrieved from Gene Expression Omnibus which the dataset named GSE20347 and derived from Search Tool for the Retrieval of Interacting Genes/Proteins which the PPI network consists of the interaction between human genes.

- GSE20347 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE20347)
- Search Tool for the Retrieval of Interacting Genes/Proteins
    - https://string-db.org/cgi/network?taskId=bZaja8QNWEYT&sessionId=b544iU0cTsPN
    - https://string-db.org/cgi/network?taskId=bizm4Ua9npug&sessionId=bQFazIXLPtDv
    - https://string-db.org/cgi/network?taskId=bDQIY5BHXwO7&sessionId=bsqROgYKpLfM
    - https://string-db.org/cgi/network?taskId=bAN6YLiXiSc0&sessionId=bsqROgYKpLfM

(d) Limitations of this study:

- Availability of high-quality data
- Difficulties in discovering relevant biomarkers.
- Computational complexity of the datasets being processed.
- Interpretation of gene clusters

1.7 Research Contribution

This research is aimed to contribute a biclustering method which able to identify potential biomarkers of esophageal cancer effectively. By developing a effective biclustring method, the accuracy and reliability of biomarker identification would be improve. This could lead to the development of effective diagnostic strategies for esophageal cancer. Since there are several biclustering methods, a few of researching will be done to make sure the method is suit to the gene expression patterns and PPI data.

1.8 Report Organization

This section explains the outline of this report.

Literature review will be included in Chapter 2. The previous study of the related research about the integration between gene expression and PPI data, the biclustering method and the biomarkers identification will be discussed in this chapter.

Chapter 3 will show the research methodology and framework used in this research in order to achieve the study.

The flowcharts and overall steps in conducting the study will be explained further in Chapter 4.

Last but not least, Chapter 5 will show and discuss the outcomes and the results. The conclusion of this study and the future work will be illustrated.

CHAPTER 2

LITERATURE REVIEW

2.1 Introduction

This chapter further discussed the details of the related research. The details of EC will be discussed further in this chapter and outlined the risk factors in causing EC. Then, the use of gene expression and PPI in discovering the biomarkers will next be explained in depth. The popular biclustering algorithms will be discussed and the advantageous and drawbacks of each algorithm will be outlined.

2.2 Esophageal Cancer (EC)

EC is the eighth most frequent cancer in the world, with over 570,000 new cases diagnosed each year (Bray et al, 2018). Since the pathophysiology of EC is less well understood than that of many other malignancies and it frequently displays an incredibly aggressive clinical picture at the time of diagnosis (Bray et al, 2018). Thus, EC is the sixth-leading cause of malignancy-related death with a 5-year survival rate ratio which is between 15-20% (Bray et al, 2018). According to Lagergren (2017), esophageal squamous cell carcinoma (ESCC) and esophageal adenocarcinoma (EAC) are the two major subtypes of EC which are proximal ESCC and distal esophageal EAC. Although ESCC is the most common pathogenic variant of EC, the incidence of ESCC and EAC varies greatly across countries and locations (Arnold, 2015). Patients with ESCC, for example, account in Asia; however, EAC is more common in Europe (Arnold, 2015).

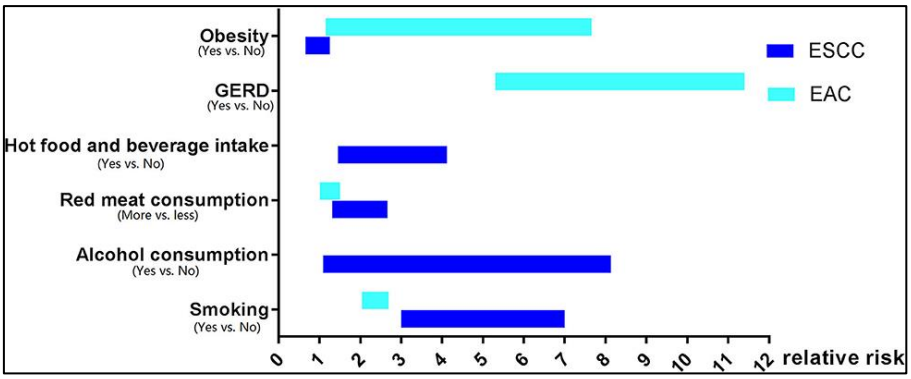


Figure 2.1: The Risk Factors for ESCC and EAC (Yang et al, 2020, p.1727).

Smoking increases the risk of developing ESCC and EAC. Unexpectedly, smoking had a greater relationship with ESCC incidence than EAC. The risk of ESCC is three to seven times higher for current smokers than it is for non-smokers. Smoking also raises the risk of EAC, however the correlation is weaker than it is for ESCC. The risk of EAC in smokers was almost two times higher.

Besides that, alcohol consumption and hot food and beverage intake only give the impact on ESCC. The risk of alcohol assumption is one to eight times higher for the drinker than non-drinker. Meanwhile, hot food and beverage intake has the risk of one to four times higher than normal people. For the people who suffer from EAC, gastroesophageal reflux disease can be one of the risk factors too.

In contrast of that, gastroesophageal reflux disease shown the effect in causing EAC but do not have correlation with ESCC. For people who suffer from gastroesophageal reflux disease have the higher chances at which five to twelve times higher to suffer from EAC.

Moreover, red meat consumption showed less effect in causing EAC than ESCC. There are one to three times higher for more red meat consumption people to get the ESCC while there is a little effect between more and less red meat consumption for EAC disease. As opposed to red meat consumption, obesity showed higher effect in causing EAC than ESCC. For obese people, the chances to diagnose EAC is one to eight times higher than normal people. Meanwhile, obesity showed a little effect between obese and normal people for ESCC disease.

2.3 Gene Expression and Protein-Protein Interaction (PPI) in Biomarker Detection

Gene expression data shows incomplete biological picture which may causing the unreliable and inaccurate result. (Karimizadeh et al, 2019). The information obtained from PPI network enable the visualization of the evolutionary links and the functional orthologous protein (Athanasios, 2017). Hence, PPI and gene expression enables the discovery of the underlying pattern on the data and obtained the reliable result.

2.3.1 Gene Expression

According to Yousef, Kumar and Bakir-Gungor (2020), extracting information from huge databases of genes that vary in expression gets difficult as high-throughput methods become advanced and massive transcriptome datasets become available. The key problem is to identify disease related information from a vast amount of redundant data and noise as gene expression data are typically limited in sample size, high in dimensionality, and noisy (Yousef, Kumar and Bakir-Gungor, 2020). Therefore, choosing the right genes and eliminating unnecessary or irrelevant genes are crucial steps in solving this issue (Yousef, Kumar and Bakir-Gungor et al, 2020). Most feature selection techniques now in use for gene expression data analysis choose genes simply based on expression values; biological knowledge is then integrated to acquire biological insights or to confirm initial findings (Yousef, Kumar and Bakir-Gungor et al, 2020).

From the understanding of Abd-Elnaby, Alfonse and Roushdy (2020), data on gene expression is a measurement of the degree of gene activity in a particular cell, tissue, or organism. Thus, it is able to provide the information for medical diagnosis as the genes in the datasets are the functional molecules that are involved in specific cellular processes (Abd-Elnaby, Alfonse and Roushdy, 2020). In summary, it is possible to obtain insight into the important underlying biological mechanisms and pathways by identifying the differential expression patterns of genes linked to a

particular disease or condition. Indirectly, these differentially expressed genes can act as potential biomarkers for a specific disease.

2.3.2 Protein-Protein Interaction (PPI)

The fundamental components of life are proteins, which are comprised of amino acids. Genes use amino acids to create peptides, which in turn create diverse proteins (Lu et al, 2020). Proteins are the building blocks of living tissue. Based on the explanation of Lu et al (2020), essential biological procedures in cells that directly affect our health, such as DNA replication, transcription, translation, and transmembrane signal transmission, depend on proteins that have specialised functions. Protein complexes, which are frequently governed by protein-protein interactions (PPIs), regulate the biological processes outlined above (Lu et al, 2020).

Cabri et al (2021) stated that PPIs are essential signalling pathways in the development of various disease states, making them ideal targets for therapeutic discovery. The role of PPIs in tumour growth is strongly correlated with protein-mediated signalling pathways that can activate numerous biological networks involved in carcinogenesis, progression, invasion, and metastasis (Cabri et al, 2021). As a result, PPI networks can be studied to find relevant proteins or nodes that function as possible biomarkers and have a significant impact on cancer pathways.

2.4 Unsupervised Clustering Machine Learning in Biomarker Detection

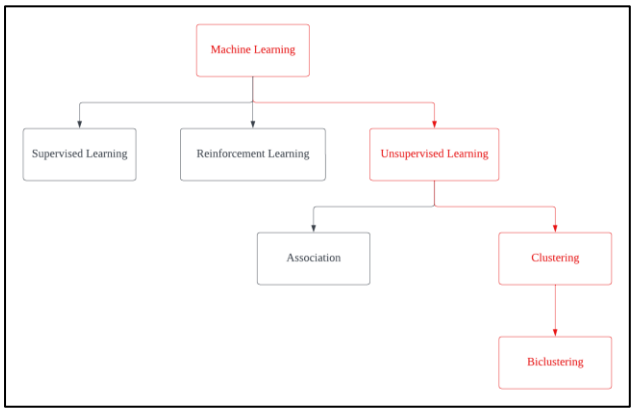


Figure 2.2: The Types of Machine Learning

Figure 2.2 illustrates the methods that are under machine learning. Our study only involves the biclustering methods which undergo the unsupervised clustering technique.

Ray (2019) proposed that a computer program is assigned to perform tasks in machine learning, and its measured performance at these tasks increases as the machine obtains more and more experience executing these jobs. Our research involves machine learning algorithms to recognize patterns and correlations in input data without providing labelled outputs. The program groups gene expression and PPI data together based on the similarity or difference by using clustering approaches. Then, the algorithm can find clusters that may influence EC by examining connections and patterns in the data. Biclustering is a technique that can be used by machine learning algorithms to iteratively assign data points to clusters while optimising a cost function that measures the similarity or distance between data points and clusters (Ray, 2019). Without explicitly providing any labelled findings, the algorithm learns to recognise patterns and correlations in the data through this iterative process (Ray, 2019). As a result, the programme can find EC biomarkers.

According to Komorowski (2022), high-dimensional datasets have been mined for hidden patterns or underlying structures using unsupervised learning due to the supervised learning requires labelling the data, which can be time-consuming and difficult. Furthermore, there could be dozens or even millions of features in high-

dimensional data, and manually labelling each data point requires a lot of resources (Komorowski, 2022). Additionally, labels for high-dimensional data cannot be readily available or be challenging to get circumstances, such as when analyzing gene expression or image data (Komorowski, 2022). Hence, using supervised learning in high-dimensional datasets could be a time-consuming project. However, without labelling the outcomes, unsupervised learning enables study of the underlying relationships and patterns in data.

Based on the research done by Wang et al (2020), unsupervised machine learning had been applied to identify the latent disease clusters and patient subgroups (Wang et al, 2020). The finding suggested that it is possible to quantify additional risk above what is expected for a particular age and gender by utilizing disease clusters to discover various potential comorbidities (Wang et al, 2020). In other words, the existence of certain co-occurring diseases raises the probability of developing a specific disease, even if the individual is of a certain age and sex (Wang et al, 2020). Thus, this data can be used to recognize high-risk groups and create more specialized preventative and treatment plans.

From the result obtained, it can be concluded that patient subgrouping based on shared traits and risks can be achieved with unsupervised machine learning techniques (Wang et al, 2020). This strategy can find relationships and patterns in patient data. Hence, by recognizing patterns and linkages in patient data, and finding distinct patient subgroups is beneficial for epidemiological analysis and research as well as enabling personalized care, which increases the effectiveness and efficiency of illness prevention, diagnosis, and treatment (Wang et al, 2020).

Based on cancer classification research done by Ayyad, Saleh and Labib (2019), the researchers proposed that using classification for gene expression data was challenging as due to the high dimensionality found in the small sample size of gene expression data (Ayyad, Saleh and Labib, 2019). Even biclustering may face to the same challenge, but biclustering can aids in addressing the multiple testing issue in the study of gene expression data, a frequent issue in classification techniques that can result in overfitting or subpar generalisation to new data (Ayyad, Saleh and Labib,

2019). Hence, classification algorithms are frequently used to group individual samples into predetermined groups based on a set of input features, but they may not be helpful for discovering new biomarkers or trends in gene expression data.

2.4.1 Biclustering Algorithm

Biclustering is frequently used in various fields of data matrix data analysis to find related entities under specific criteria (Liu et al, 2020). According to Gu and Liu (2008), biclustering of gene expression data looks for regional patterns of gene expression and biclustering of PPI network is aimed to identify the subsets of interacting protein. Based on the finding of Eren et al (2013), performance of the algorithms is different based on the bicluster model chosen. It is crucial to take into consideration the pattern of the data and choose the correct parameters for each method (Eren et al, 2013). Hence, the most common algorithms will be studied to identify the suitable method used for the study.

2.4.1.1 Correlated Pattern Biclustering (CPB)

CPB is a biclustering technique which is used for finding clusters of genes linked to some target genes of interest (Eren, 2012). According to the finding of Eren (2012) and Yun and Yi (2013), CPB is predicted to do well on both constants and the upregulated bicluster model in model experiment to test whether the algorithm can give complete and perfect result. However, CPB recovery decreases as the upregulated bicluster model rises as increased levels of differential expression make it harder to identify underlying patterns of association between genes (Yun and Yi, 2013). This behavior makes logical because CPB finds biclusters with high row correlations, which means CPB is useful for identifying co-expression genes (Yun and Yi, 2013).

Besides that, Eren (2012) stated that CPB is highly sensitive to noise which lowers the accuracy of algorithm findings and causes false positive identifications. For the number experiment, CPB showed little effect on the result (Eren, 2012). The

finding of Yun and Yi in the overlapping experiment for CPB model showed that, the capacity of CPB to recover biclusters declines as the amount of overlap between biclusters declines (Yun and Yi, 2013). For your information, number experiment referred to the number of biclusters used for the experiment while overlap experiment referred to the overlapping with two biclusters by different amounts of overlapping elements in rows and columns.

In conclusion, CPB is performed better even the large numbers of biclusters is used and the data show higher correlation between rows and columns. In contrast. CPB had the limitations which are sensitivity to noise and low ability to detect the bicluster that there is highly differential expression. Due to these characteristics, CPB is not suitable for identifying the biomarkers of esophageal cancer as the datasets used needed to detect gene clusters that exhibit differential expression when compared to normal tissues.

2.4.1.2 QUBIC

QUBIC is a biclustering technique used in data analysis to discover sets of genes or traits that display coordinated behaviour which are the genes that work together to carry out specific functions such as metabolic pathway across a set of conditions or samples (Renc et al, 2021). Biclustering techniques cluster rows and columns of a dataset concurrently, and QUBIC uses a Bayesian framework to locate subsets of rows and columns with comparable behaviour (Renc et al, 2021). Renc et al (2021) had carried out the running experiment to test the time taken for QUBIC algorithm to complete the bicluster task based on the given datasets. The results showed QUBIC able to run faster to perform the bicluster of datasets (Renc et al, 2021). However, Xie et al (2020) stated that QUBIC would be time consuming if large datasets had been applied to the algorithm (Xie et al, 2020).

According to the study done by Cui et al (2020), the performance of QUBIC had been evaluated by using different sets of datasets. The results showed QUBIC had low performance on the experiment. The experiment showed that QUBIC algorithm

had lower average volume of the biclusters found and average correlation coefficient within a bicluster. However, QUBIC had the highest average mean squared residue and the average connectivity value, which measures the average number of other biclusters with a bicluster is connected to when compared to Cheng & Church (CC) algorithm and the proposed algorithm.

As a final point, QUBIC had the better execution time for biclustering the datasets. However, when QUBIC applied to the large datasets, the execution time would be slower. Besides that, the higher average mean square residue and higher average connectivity value indicates that QUBIC had low accurate and reliable result.

2.4.1.3 Bayesian Biclustering (BBC)

The Bayesian Biclustering (BBC) algorithm automatically groups the rows and columns of a dataset into "Checkerboard" clusters that are exhaustive and exclusive (Pinto, Gates and Wang, 2020). Pinto, Gates and Wang (2020) conducted studies that evaluated the performance of BBC under various conditions.

Different degrees of noise were applied to the dataset by Pinto, Gates and Wang (2020). According to experimental findings, the biclustering algorithm's accuracy declines as noise level rises. Due to the noise, which makes it challenging for the algorithm to recognize bicluster patterns and indirectly causing the performance accuracy of the BBC algorithm to decrease. Additionally, Pinto, Gates and Wang (2020) demonstrate that the BBC algorithm takes longer time to run when large datasets are used. Meeds and Roweis, S (2007) proposed that BBC is a biclustering algorithm which robust to missing values. Hence, we can conclude that BBC able to produce an accurate and meaningful results even there are missing values in the datasets.

However, Do, Muller and Tang (2005) indicated that with the help of Markov chain Monte Carlo (MCMC) algorithms, bayesian algorithm can deal with missing data and estimate the posterior probability distribution of unknown parameters given

observed data and missing data. However, the degree and pattern of missingness can all have an impact on how successfully Bayesian approaches handle missing data. The accuracy and reliability of Bayesian approaches may be compromised if there is an extensive amount of missing data (Do, Muller and Tang, 2005).

Taking everything into account, BBC algorithm perform well in lower level of noise, and has shorter execution time in evaluating small datasets. Besides that, BBC algorithm able to produce accurate and meaningful result even there are missing values in the datasets. Nevertheless, the existence of many missing values in a dataset might result in overfitting and false positives in analysis findings.

2.4.1.4 Binary inclusion-Maximal (BiMax)

BiMax is a simple reference technique that locates biclusters of 1s in a binary matrix (Eren, 2012). It uses a divide and conquer strategy to iteratively bicluster the data matrix (Eren, 2012). The BiMax algorithm searches a matrix for submatrices with only 1s in it (Eren, 2012). These sub-matrices are viewed as possible biclusters, and the algorithm builds these potential biclusters iteratively by including rows and columns that have a lot of 1s in common (Eren, 2012). When no additional rows or columns can be added 1s in the bicluster, the growing process comes to an end (Eren, 2012). This results in a collection of biclusters with a high co-occurrence rate of 1.

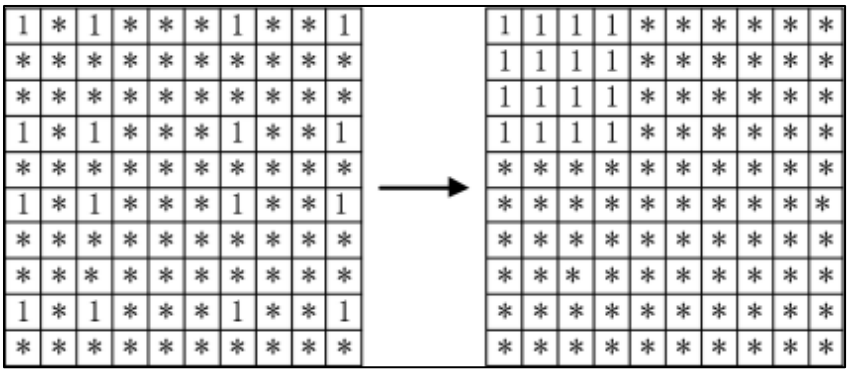


Figure 2.3: Biclusters of 1's in a Binary Matrix

According to the study done by Bustamam et al (2020), the BiMax algorithm works well in clustering protein-protein interactions, particularly for binary data

compare to local search framework based on pairs operation and LCM-MBC. BiMax is the best approach for classifying binary protein-protein interaction data, as demonstrated by the experiment conducted by Bustamam et al (2020) in identifying the bicluster on interacting proteins between HIV-1 and humans. Despite that, Voggenreiter, Bleuler and Gruissem (2012) believed that the BiMax method would work best with input data that was limited in size. BiMax took longer time to process large sample size.

Furthermore, Castanho, Aidos and Madeira (2020) indicated that BiMax is useful and highly quick algorithm capable of detecting simple structures. The BiMax technique has the drawback of only looking for binary biclusters, which restricts its capacity to locate useful biclusters in the dataset (Castanh, Aidos and Madeira, 2020). This is because discretizing data into binary form is a very particular procedure that is unable to account for all possible ranges of values in the data (Castanho, Aidos and Madeira, 2020). Therefore, when the approach is applied to datasets that do not fit binary bicluster models well, bad results may be obtained (Castanho, Aidos and Madeira, 2020).

Last but not least, BiMax is very effective at detecting simple structures in binary data. Additionally, BiMax has been demonstrated to function better with fewer samples. When the dataset contains continuous data that cannot be transformed into discrete values, BiMax performs worse as well as finds fewer relevant biclusters on larger datasets.

2.4.1.5 Plaid

The value of a certain element is determined by the plaid model's calculation of a particular submatrix for each cell; this value can be interpreted as the number of contributions generated by a specific bicluster (Siswantining et al, 2021). According to the statement made by Siswantining et al (2021), each component of the matrix in a plaid model indicates the contribution of a certain bicluster to the overall level of gene expression under a specific circumstance. To be illustrated, the plaid model breaks

down the original matrix of gene expression data into a new matrix that demonstrated the contribution of a certain bicluster to the overall level of gene expression.

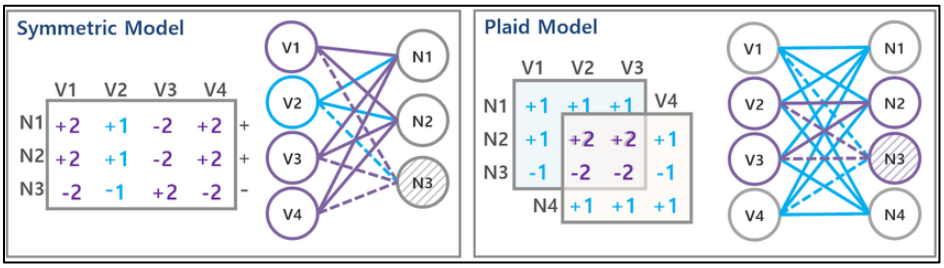


Figure 2.4: The Working Theory of Plaid Biclustering Model (Henriques and Madeira, 2015, pp 1-15)

The plaid model's ability to simulate biclusters that may overlap in order to obtain the correct model is one of its strengths (Siswantining et al, 2021). The plaid model enables it to capture more complex patterns in the data than typical bicluster approaches that assume non-overlapping biclusters. This enables a more precise and thorough depiction of the data's underlying structure. The experimental findings and analysis lead Siswantining et al (2021) to the conclusion that low coherence variance colon cancer data can be analysed using bicluster analysis on the plaid model. Low coherent variance in plaid models could be a sign that the model accurately captures data patterns.

According to Karim, Kanaya and Altaf (2019), spectral and plaid biclustering model achieved second highest in the performance of average cluster relevance compared to the proposed algorithm by Karim, Kanaya and Altaf et al (2019) which has the highest performance of average cluster relevance. For your information, the average cluster correlation metric assesses how successfully the biclustering method detects related biclusters in data.

Kocatürk, Altunkaynak and Homaida (2019) conducted an experiment to compare the quality of biclustering algorithms using data envelopment analysis methods. Data envelopment analysis can assist to select the most effective parameters for several algorithms and ranking them according to specified criteria (Kocatürk, Altunkaynak and Homaida, 2019). Based on the results obtained, plaid model obtained an overall good performance compared to others biclustering algorithm. In a nutshell,

plaid biclustering model advanced in capturing overlapping biclusters and able to performance better than other biclustering algorithms.

2.4.1.6 Iterative Signature Algorithm (ISA)

Iterative Signature Algorithm (ISA) is a biclustering algorithm that can generate overlapping biclusters (Freitas et al, 2011). ISA produces good outcomes on a number of synthetic and real-world datasets (Freitas et al, 2011).

According to Freitas et al. (2011), codon-pair context maps of sequenced genomes could use the ISA algorithm approach. ISA can find hidden homogenous groups, however errors and outliers in the dataset can have a big impact on the mean's ability to quantify centrality (Freitas et al, 2011). The usage pattern of the set of codon pair can be summed up using the average of the biclusters as a measure of centrality (Freitas et al, 2011). The means of the biclusters, however, can be greatly altered and may not accurately indicate group centrality if there are errors or outliers in the data set that affect the correlation between rows and columns (Freitas et al, 2011). Before executing ISA under such circumstances, it might be required to employ additional measures or eliminate mistakes and outliers (Freitas et al, 2011).

However, Supper et al (2007) proposed that a well-known issue with ISA is that they favour strong signals. The ISA algorithm frequently prioritises strong signals in the data and may overlook weaker signals or patterns that may be significant but are less evident (Supper et al, 2007). As a result, the method may detect incomplete or biased biclustering findings. Furthermore, the experiment done by Sutheeworapong et al (2012) indicates that ISA algorithm had lower gene coverage and gene overlap. Greater gene coverage is often regarded as preferable because it indicates that more genes are being examined, leading to a greater understanding of the biological system being investigated (Sutheeworapong et al, 2012). Higher gene overlap may be a sign that biclusters are capturing more widespread gene expression patterns that are shared by a variety of biological processes. Hence, ISA may not be useful for investigating

datasets with a lot of weak signals or for identifying double clusters with limited gene overlap.

2.4.1.7 Spectral

A data matrix with a checkerboard structure, which can be thought of as a composition of constant biclusters in a single matrix, can be used to illustrate the goal of spectral biclustering algorithms: to discover subsets of characteristics and conditions (Shaharudin et al, 2019). The technique effectively recognizes these checkerboard arrangements even when the underlying biclusters are not precisely aligned using a spectral clustering approach (Shaharudin et al, 2019). As a result, it may be used to analyze high-dimensional datasets such gene expression data.

Bicluster visualization was tested in a research study by Liu et al (2022). The outcomes reveal that when the biclusters are small and the noise level is low, the spectral biclustering method recovers the real patterns with excellent accuracy. Spectral biclustering is an effective technique for identifying unique molecular subtypes in patient populations based on gene expression profiles (Liu et al, 2022). By clustering patients based on gene expression patterns, spectral biclustering can identify important gene expression patterns that may be related with varied illness outcomes or treatment responses. The results for patients can be improved by using this data to create more precise prognostic models and better risk stratification techniques (Liu et al, 2022).

To compare three or more related groups to see if there are any significant differences between them, the nonparametric Friedman test is a statistical test (Branders, Schaus and Dupont, 2019). It functions by ranking the observations inside each group and comparing the average ranking between groups. If the mean ranks differ significantly between the groups, there are significant variations between them (Branders, Schaus and Dupont, 2019). The authors compared biclustering algorithms using a nonparametric Friedman test. The methods under examination are graded based on the number of enriched biclusters they produce for each dataset. The result

showed spectral biclustering able to obtain higher enrichment analysis. In conclusion, spectral biclustering method effective when allocate to the lower noise level with small data and able to present greater enrichment analysis.

2.4.1.8 Order Preserving Submatrix (OPSM)

The OPSM is a continuous bicluster that is monotonically increasing or decreasing with the degree of gene expression (Maind and Raut, 2019). In other words, biclusters that exhibit repeated patterns of increased or decreased expression levels across genes and samples are identified using OPSM. A selection of genes that are co-regulated under a subset of circumstances are referred to as having a consistent pattern (Maind and Raut, 2019).

Research by Maind and Raut (2019) on column subspace extraction and pattern recognition demonstrates that OPSM can accurately extract biclusters, extract biclusters that overlap, and provide stable output. In order to extract the column subspace, a subset of the original data matrix's columns must be chosen, and in order to extract patterns from the column subspace, biclusters must be located inside these chosen columns (Maind and Raut, 2019). This method provides additional flexibility in discovering biclusters because it identifies biclusters that do not always span all rows and columns of the original matrix (Maind and Raut, 2019). However, the performance results of OPSM on synthetic data of column coherent evolution are unsatisfactory (Maind and Raut, 2019). Column coherent evolution describes a situation in which samples may be divided into groups and columns (genes) are highly connected within each group.

To compare methods, which frequently provide inadequate or misleading information on a single model, each bicluster was evaluated on a synthetic dataset (Eren et al, 2013). It turns out that OPSMs do not filter their output, which causes them to produce a large number of incorrect biclusters and lower their correlation scores. (Eren et al, 2013) Li (2020) proposed that OPSM cannot adequately analyse gene expression datasets.

2.4.1.9 Cheng & Church (CC)

Cheng and Church (CC) were the first to propose biclustering for finding genomes that may overlap and/or exhibit high similarity in gene expression data matrices (Di Iorio, Chiaromonte and Cremona, 2020). Finding the bicluster that maximises the score function while considering specific constraints is the objective of the biclustering issue as it is formulated by the CC algorithm framework. (Tanay, Sharan and Shamir, 2005) In most cases, the similarity of their gene expression patterns conditional on a subset is used by the scoring function to determine the quality of candidate biclusters. (Tanay, Sharan and Shamir, 2005) These restrictions guarantee that the discovered biclusters have a particular dimension, form, or structure (Tanay, Sharan and Shamir, 2005). The CC algorithm employs a heuristic search approach to quickly explore the space of potential biclusters and find biclusters that match the requirements and optimise the score function (Tanay, Sharan and Shamir, 2005).

According to Yang et al (2003), the CC technique is recognised to have limitations in discovering big biclusters with high consistency in noisy datasets. The initialization and ordering of the rows and columns in the data matrix have an impact on the greedy approach of the algorithm (Yang et al, 2003). This means that the output of an algorithm can depend on how the data was initially sorted and processed, and even tiny changes in the row- and column-order can have a significant impact on the final product. Yang et al (2003) also proposed that the bicluster discovered may not be the ideal bicluster since the CC technique is vulnerable to local optima. As the CC algorithm discovers more biclusters, it replaces them with random data, making it more difficult to find larger, more coherent biclusters (Yang et al, 2003).

According to Eren et al (2013), CC have long run times if the settings are not set properly. CC were successful in identifying a significant number of abundant biclusters in gene expression data (Eren et al, 2013). Abundant double clusters, however, might not be as trustworthy or biologically significant (Eren et al, 2013) .Enriched biclustering enables a more thorough comprehension of gene expression patterns and their relationship to biological processes, enabling a more in-depth comprehension of underlying mechanisms (Eren et al, 2013).

2.5 Summarizing The Biclustering Methods

According to the literature review that had been done, most of the biclustering algorithms have limitations to the higher noise level and sample size.

Table 2.1: Summarize the Biclustering Algorithms

Biclustering Algorithms	Advantages	Disadvantages	Citation
CPB	• Work well in synthetic datasets • Perform well in large numbers of biclusters	• Sensitive to noise • Low ability to detect higher differential expression	• Eren (2012) • Yun and Yi (2013)
QUBIC	• Better execution time	• Low accurate and reliable result	• Renc et al (2021) • Xie et al (2020)
BBC	• Well-handled missing values	• Sensitive to noise level and size	• Pinto et al. (2020) • Meeds and Roweis, S (2007)
BiMax	• Effective for simple structure	• Sensitive to size • Limited to discrete values datasets	• Voggenreiter et al (2012) • Castanho et al (2020)

Biclustering Algorithms	Advantages	Disadvantages	Citation
Plaid	• Advanced in capturing overlapped bicluster • Low coherent variance	• Sensitive to parameters used	• Siswantining et al (2021) • Karim et al (2019) • Kocatürk et al (2019)
ISA	• Able to find hidden homogenous group	• Sensitive to errors and outliers • Favor strong signals	• Freitas et al. (2011) • Supper et al (2007)
Spectral	• Able to identify unique molecular subtypes • Higher enrichment analysis	• Sensitive to noise level and sample size	• Liu et al (2022) • Branders et al. (2019)
OPSM	• Extract overlapped bicluster accurately • Provide stable output	• Do not filter output • unable to analyse gene expression datasets	• Maind and Raut (2019) • Eren et al (2013)
CC	• Able to identify large number of bicluster	• Performance limited to higher noise level • Vulnerable to local optima • Long execution time	• Yang et al (2003) • Eren et al (2013)

2.6 Classification Methods for Gene Expression Data

For the past few years, scientists have been exploring through vast volumes of gene expression to extract useful knowledge that can help categorize cancers (Ayyad, Saleh and Labib, 2019). Among the classification methods used are Support Vector Machine (SVM), K-Nearest Neighbours (kNN), neural networks and decision trees. Hence, the review on the classification methods for identifying potential biomarkers from gene expression data will be focused on these four methods.

SVM is a well-liked technique for both linear and nonlinear classification (Uddin et al, 2019). According to Uddin et al (2019), kNN is a nonparametric technique that determines the class of a new observation based on the k-nearest neighbours' predominant class. Meanwhile, neural networks are algorithms that are modelled after the structure and functioning of neural networks in the human brain. These algorithms can learn from information, identify patterns, and make predictions or categorizations (Uddin et al, 2019). A decision tree is a tree-based machine learning technique composed of nodes and edges used to explain the data separation or classification process in which begins from the starting point till an outcome is produced (Charbuty and Abdulazeez, 2021).

2.6.1 Support Vector Machine (SVM)

According to Steardo et al (2020), SVM has demonstrated outstanding results in precisely and accurately diagnosing people with schizophrenia. As the most well-known and well-established machine learning technology, it is frequently used as a standard to measure other methods against. SVM is flexible as it can handle classification and regression tasks (Steardo et al, 2020). However, it should be emphasised that SVM implementation can be expensive and complexity (Steardo et al, 2020).

While doing the research on the discovery of biomarker for cancer gene expression data, researchers found that SVM's ability to handle high-dimensional

datasets, particularly when the sample size is small compared to the number of features, is one of the benefits of employing it to classify microarray gene expression profiles (Almugren and Alshamlan, 2019). However, SVMs require a lot of processing power, especially when working with large datasets or complex models (Almugren and Alshamlan, 2019).

2.6.2 K-Nearest Neighbours (kNN)

Based on the research of the information from heart disease prediction done by Uddin et al (2019), kNN can quickly classifies instances and is simple to understand. Second, it is adaptable to noisy data and capable of handling situations with missing attribute values (Uddin et al, 2019). Finally, kNN is flexible and can be utilised for both classification and regression tasks (Uddin et al, 2019). However, the number of neighbours (k) and the distance metric utilised, which are crucial factors in its implementation, might have an impact on the performance of kNN (Uddin et al, 2019).

Besides that, kNN algorithm has drawbacks (Uddin et al, 2019). The kNN algorithm is computationally expensive as the number of attributes rises. This is because kNN need to calculate the distance between the attributes (Uddin et al, 2019). Furthermore, kNN treats all attributes equally which may consider the less important features and lacking information about the importance of attributes for effective classification (Uddin et al, 2019).

2.6.3 Neural Networks

Artificial neural networks can capture and simulate complex relationships that might exist between variables (Uddin et al, 2019). This makes them outstand for situations where the underlying patterns are inherently nonlinear, allowing them to identify complex patterns and make accurate predictions (Uddin et al, 2019). Artificial neural networks (ANN) are flexible and can perform both classification and regression tasks (Uddin et al, 2019).

Artificial neural networks frequently function as "black box" models, which means that it is difficult to understand or describe exactly how they make decisions (Uddin et al, 2019). It is challenging to comprehend why the network produced a specific prediction because of this lack of openness. Moreover, training artificial neural networks for complex classification tasks or massive volumes of data may be computational expensive and time-consuming (Uddin et al, 2019).

2.6.4 Decision Trees

Decision trees have difficulty in gene expression data since there are many more features than observations (Czajkowski and Kretowski, 2019). Even though learning algorithms may discover splits that precisely divide the training data, these splits frequently correspond to noise rather than important patterns (Czajkowski and Kretowski, 2019). As a result, decision tree techniques frequently result in uncomplicated trees that successfully identify previously unseen examples but perform poorly when applied to the data that the model has not been encountered before (Czajkowski and Kretowski, 2019).

The decision trees produced a hierarchical structure which is simple to visualize and analyze, which is helpful for outlining the decision-making process (Uddin et al, 2019). Second, because decision tree algorithms can handle various types of data, including numerical, nominal, and categorical data, it typically requires less data preparation than other algorithms (Uddin et al, 2019). Decision trees have the potential to achieve high predictive accuracy by efficiently partitioning the feature space based on available data (Uddin et al, 2019).

2.7 Summarizing The Classification Methods

Table 2.2: Summarized the Selected Classification Methods

Classification Methods	Advantageous	Disadvantageous	Citation
SVM	• flexible • handle high-dimensional datasets	• can be expensive and complexity. • require a lot of processing power	• Steardo et al, 2020 • Almugren and Alshamlan, 2019
kNN	• simple and adaptable to noisy data • capable of handling situations with missing attribute values.	• Performance based on parameter. • computationally expensive • treats all attributes equally	• Uddin et al, 2019
Neural Network	• can capture complex relationships. • flexible	• difficulty visualizing the decision-making process. • time consuming	• Uddin et al, 2019
Decision Tree	• simple to visualize and analyze. • requires less data preparation. • have the potential to achieve high predictive accuracy	• difficulty in gene expression data • splits frequently correspond to noise rather than important patterns.	• Czajkowski and Kretowski, 2019 • Uddin et al, 2019

2.8 Identifying Optimum Number of Cluster

The purpose of clustering is to arrange data points into groups in which the cluster members are as similar as feasible, and the cluster between clusters are as distinct as possible (Hayasaka, 2022). This indicates that under optimal clustering, variation within clusters is low while variation across clusters is high.

The quality metric for the calculation of number of clusters are inertia and silhouette coefficient (Hayasaka, 2022). Inertia quality metric entails calculating the sum of squared distances between data points and the centres of each cluster meanwhile silhouette coefficient seeks to aggregate variation within and between clusters (Hayasaka, 2022). Among of the approaches to obtain the optimum number of clusters are elbow method, silhouette method and gap statistic (Hayasaka, 2022).

According to the experiment done by Hayasaka (2022), the interpretation of elbow plots is sometimes subjective, the silhouette coefficient and gap statistical approaches can correctly quantify the number of clusters. Gap statistics, however, include computations that could not always provide the same result (Hayasaka, 2022).

According to Kumar (2021), the elements that each technique considers while assessing the quality of clustering are the fundamental distinction between elbow method and silhouette score. While silhouette scores consider other factors including variance, skewness, and value differences, elbow approaches primarily concentrate on determining Euclidean distances (Kumar, 2021).

Elbow Method uses an approach that is clear and straightforward (Kumar, 2021). Furthermore, the Elbow method is an effective computing method that doesn't need a lot of calculations or iterations (Kumar, 2021). Kumar (2021) also stated that, If the sum of square error line graph forms an arm, then the Elbow Method is the suitable method for the finding of optimum number of clusters. Hence, the Elbow Method will be used for this research. This is because a clear “elbow” diagram was able to be obtained from the datasets.

2.9 Chapter Summary

Biclustering approaches have been studied to find the best approach for assessing gene expression data and PPI networks. After consideration, the plaid model was chosen as the biclustering technique to identify potential esophageal cancer biomarkers. The ability of the plaid algorithm to analyze overlapping biclusters using a matrix factorization method that allows row and column clusters to overlap in order to reveal deeper and more complete biclusters. Plaid is a suitable method for studying gene expression data and finding biomarkers because it can provide a more comprehensive understanding of the underlying biological processes. Furthermore, plaid shows low coherence variance, which suggests that the gene expression levels inside the biclusters are strongly correlated and have minimal volatility. This is advantageous for biclusters because it demonstrates that there is a high correlation between genes and circumstances in biclusters, increasing the probability that they represent biologically significant groups. In other words, low coherence variance suggests functional relationships between genes within biclusters and probably shared biological functions. Research methodology will be discussed in the next chapter.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The research framework is covered in this chapter. A research framework is crucial because it provides authors a clear road map and makes sure that any relevant problems are taken into account and handled. There will be four phases to the entire study. The entire process, from the planning of the study to the verification of the findings, will be clearly explained. The datasets chosen for the study, the performance measurements employed to calculate the approaches' performance, and the hardware and software requirements will all be clarified in this chapter.

3.2 Research Framework

A few phases were carried out to ensure full adherence to the study protocol in order to accurately identify and gather possible biomarkers for esophageal cancer.

Research planning and initial study will be covered in phase one. To determine the issue as well as the objectives and goals of the research, a review of the relevant literature will be conducted during this stage. The data gathered will next be preprocessed and normalized. The second stage will go through how the plaid model can bicluster the input data to find possible biomarkers. The third phase will classify potential biomarkers and determine performance accuracy. The chosen biomarkers will then be validated by the biological knowledge base in a fourth phase to make sure they are susceptible to esophageal cancer.

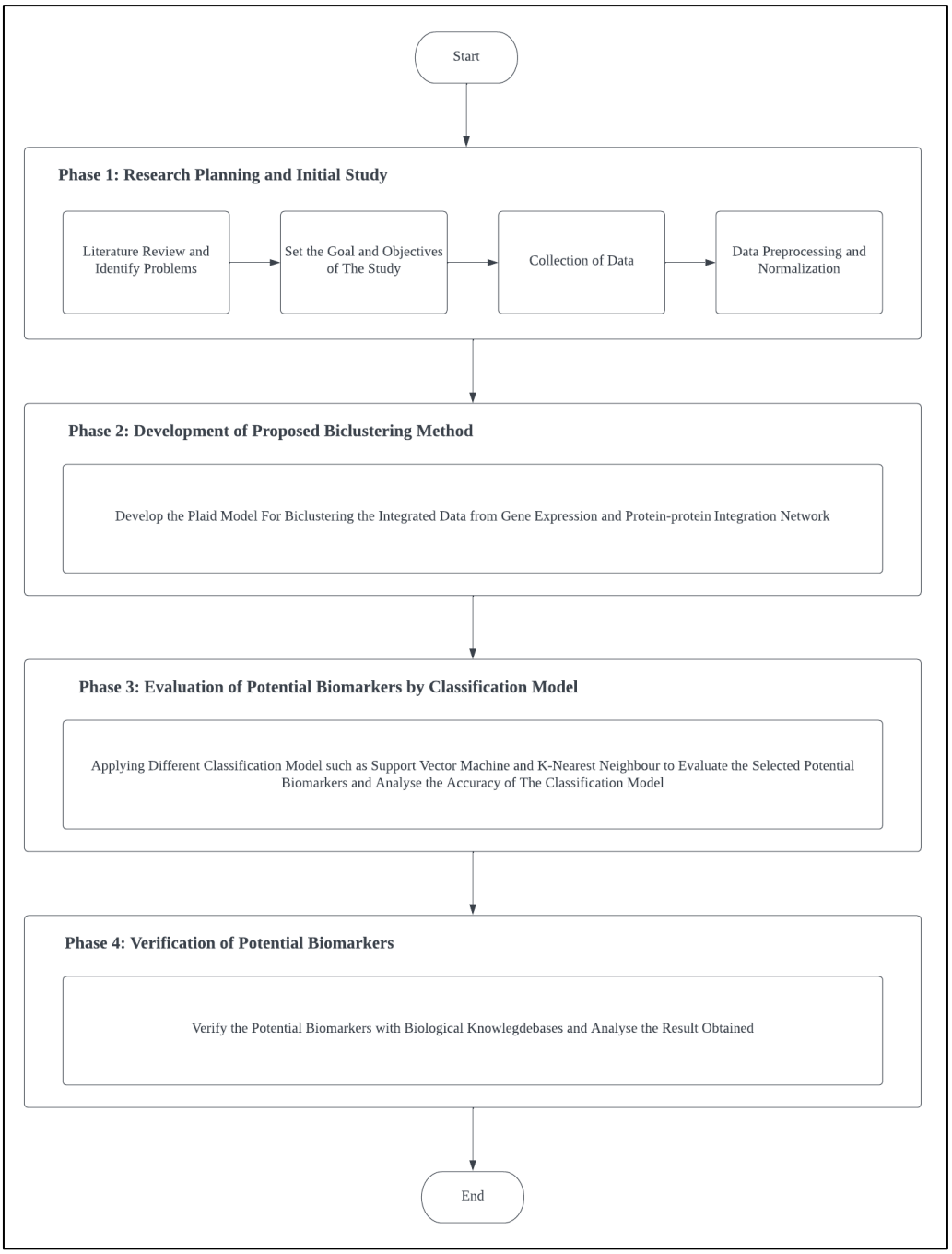


Figure 3.1: Research Framework

3.2.1 Phase 1: Research Planning and Initial Study

To make sure a study is viable, relevant, and solves a key research issue, it is essential to conduct preliminary research and develop a research plan before starting. Authors can determine the appropriate plan of study, methodology, data collecting, and analysis procedures needed to accomplish their research aims by conducting

adequate preparation and exploratory research. Researchers can improve their chances of success and decrease the risk that they will waste time on ineffective or unrelated research issues by carefully preparing and conducting preliminary research.

Figure 3.1 shows that four activities are necessary to carry out for future work. Literature review, problem identification, defining the study's goals and objectives, collecting data for the study's input, and preparing and normalizing the data are all tasks that fall within the first phase. A literature review is conducted initially since it is a crucial step in the study's process, and it allows the author to become familiar with the topic that desires to explore further. By identifying the appropriate methods and techniques used by other researchers in similar studies, a thorough literature review aids in preparing for and carrying out of the author's own research and helps authors prevent duplicating previous studies. The author can get a complete view of problem areas by analyzing previous issues-related work done by other researchers. This allows the author to strategically plan their study. Thus, goals and objectives can be defined.

In research, data collecting is critical because it acts as the foundation for analysis and interpretation. Without effective data collection, research findings could be inaccurate or misleading, and the stated aims of the study would not be met. Data collection involves finding relevant data sources, choosing appropriate data collection techniques, and ensuring the accuracy and precision of gathered data. In the context of gene expression and PPI network analysis, data collecting involves gathering gene expression data or PPI network data from relevant databases or experimental studies and ensuring that the data are of high quality as well as relevant to the research subject under consideration.

Hence, there are two datasets chosen for this study. One of the datasets was obtained from GEO database which is named GSE20347 while another dataset was obtained from STRING websites which consists of the human genes. The details of the datasets had been further discussed under 3.3.

3.2.2 Phase 2: Development of Proposed Biclustering Method

Data clustering analysis works to group variables in a data matrix according to a certain global pattern, signifying a pattern generated in rows or columns to be considered. Bicluster analysis, in contrast to cluster analysis, seeks to identify regional patterns in huge data matrices (Siswantining et al, 2021). According to Siswantining et al (2021), plaid modelling is a biclustering technique that sums the values given by many overlapped biclusters to indicate the value of each element in a data matrix. In contrast to traditional biclustering techniques, which only support binary biclusters (whether a data point is a member of a bicluster or not), plaid models allow data points to be members of numerous biclusters with various intensities. This enables plaid models to capture complex patterns in the data, such as biclusters that overlap or have varied sizes.

The Plaid model can be thought of as a method of breaking down the original data matrix into a collection of biclusters, each of which represents a distinct pattern in the data, and then using these patterns to reconstruct the matrix. The rebuilt matrix can be used to visualise the relationships between various patterns and to identify the genes or traits that each pattern most closely resembles.

The general flow of the plaid model had been further discussed in 3.4.

3.2.3 Phase 3: Evaluation of Potential Biomarkers by Classification Models

With the use of data mining, classification is a machine learning technique that identifies higher-level and more advanced information by predicting and/or classifying data into specified classes or groupings (Otchere et al, 2021). Based on the finding of literature review, SVM will be applied to the selected potential biomarkers and the accuracy of performance will be calculated by confusion matrix.

3.2.4 Phase 4: Verification of Potential Biomarkers

During the classification process, the chosen biomarkers undergo training and testing. The biological knowledge base will subsequently be used to validate the biomarkers with the highest accuracy. Biological knowledge bases are enormous collections of biological data, including gene sequences, protein activities, pathways, and disease connections. Gene sequences, protein structures, and functional descriptions can be found in the NCBI and UniProt. Researchers can search for gene or protein sequences linked to potential biomarkers and check the expression levels of these biomarkers in various tissues or cells linked to the disease to validate biomarkers for a specific disease. In general, combining biomarker data with biological knowledge bases can offer insightful information about the underlying biology of a disease or biological process and aid in the identification of prospective targets for drug development and personalized therapy.

3.3 Datasets

Two datasets were applied in this research. One of the datasets obtained from Gene Expression Omnibus (GEO), which is named GSE20347. It is data of gene expression in esophageal cancer. GSE20347 consists of 34 samples, where 17 of them are tumors and 17 of them are normal. The dataset illustrated the gene expression values of the samples. The gene symbol (red color box) showed the genes are involved in the development of esophageal cancer. The GSM (blue color box) is the sample.

	Gene Symbol	GSM509787_E1507N.CEL	GSM509788_E1520N.CEL	GSM509789_E1521N.CEL
0	DDR1 /// MIR4640	10.414177	10.250918	10.046812
1	RFC2	6.839942	6.511217	6.683490
2	HSPA6	4.752045	5.115767	5.040198
3	PAX8	7.561694	7.953933	7.900248
4	GUCA1A	3.596421	3.603976	3.435885
...	...	...	...	...
22272	NaN	5.787397	5.913325	4.450484
22273	NaN	7.330281	7.202484	4.830335
22274	NaN	3.363339	3.409699	3.294732
22275	NaN	3.768794	3.853740	3.716894
22276	NaN	3.479989	3.491986	3.473315

22277 rows × 35 columns

Figure 3.2: Gene Expression Data of the GSE20347

Meanwhile, another dataset obtained from Search Tool for the Retrieval of Interacting Genes/Proteins (STRING), illustrated the PPI network of human genes. There are four different databases, Reactome, KEGG, DISEASES and Monarch presented in the STRING for the PPI human disease network. Hence, the data of all databases had been used for further interpretation. There was a total of 3506 genes that were visible in the PPI network. Both data can be obtained from the link given respectively under Chapter 1. There are nine types of evidence used in STRING to calculate the score for the PPI network, which are neighborhood on chromosome, gene fusion, phylogenetic cooccurrence, homology, coexpression, experimentally determined interaction, database annotated and automated textmining. Nine types of evidence will then be calculated for the combined score. Node 1 and node 2 (red color box) showed the genes that are interacting while the combined score (blue color box) indicated the evidence score of how likely two genes are interacted with.

	#node1	node2	combined_score
0	AAAS	VIP	0.444
1	AAAS	MC2R	0.463
2	AAAS	LIG1	0.497
3	AAAS	POMC	0.566
4	AAAS	POM121	0.600
...	...	...	...
14629	ZNF644	LRPAP1	0.549
14630	ZNF644	SCO2	0.561
14631	ZNF644	P4HA2	0.602
14632	ZNF670	PIGW	0.400
14633	ZNF670	OPTN	0.408

Figure 3.3: The PPI Network of the Human Genes that Showed in Tabular Form

Table 3.1: Features Description of PPI Network

Features	Description
Node 1, Node 2	Proteins In the Network
Node 1 String ID, Node 2 String ID	Unique Identifier for The Proteins
Neighborhood on Chromosome	The probability that two proteins have similar functions if their genes are located adjacent to one another in the genome.
Gene Fusion	The probability that two proteins are functionally linked if they are encoded by the same gene that has been fused in a different organism.
Phylogenetic Cooccurrence	The possibility that two proteins are functionally connected if their genes co-occur in different genomes.
Homology	If two proteins show significant sequence similarity across several species, they have the potential to have similar activities or engage in similar biological processes.
Coexpression	The probability that two proteins are connected functionally if their genes are expressed in many samples.
Experimentally Determined Interaction	The probability that two proteins are connected functionally if high-throughput studies demonstrate their physical interaction.
Database Annotated	A confidence score provided to an interaction based on its presence in other biological databases.
Automated Textmining	If two proteins are discussed together in the scientific literature, the probability that they are functionally connected increases.
Combined Score	A confidence score for the interaction of two proteins based on the combination of nine types of evidence.

3.4 The General Flow of Plaid Model

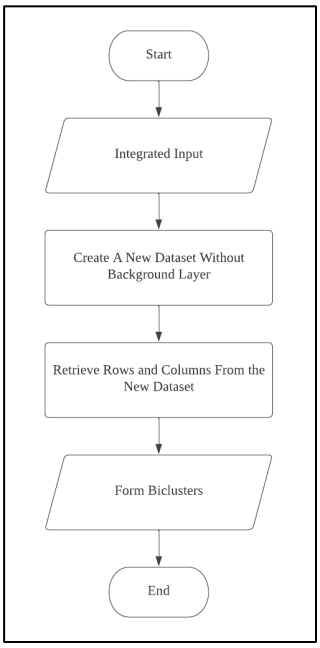


Figure 3.4: General Flow of Plaid Model

The basic concept of the plaid model is it formed a new layer based on the input data that had done the data preparation step. This new layer will exclude the background layer that accountable in the input data. Then, the algorithm repeatedly retrieves the rows and columns randomly from the new layer until the maximum number of the biclusters had formed. For an example, when the maximum number of the biclusters is four, then the plaid algorithm will keep retrieve the rows and columns from the layer created. When there are four groups of biclusters is formed, then the process will be terminated.

3.5 Performance Measurement

Confusion matrix will be used in this research to calculate the accuracy of the classification model. Then, biological context verification will be used to verify the selected biomarker. Sum of Square Method will be used in the Elbow Method to find the optimum number of biclusters.

3.5.1 Confusion Matrix

A confusion matrix is a table that compares the predicted classes in a test dataset to the actual classes in order to evaluate the effectiveness of a classification algorithm. The number of true positive, true negative, false positive and false negative predictions are indicated (Luque et al, 2019).

Table 3.2: Confusion Matrix

	Predictive Positive	Predictive Negative
Actual Positive	True Positive (TP)	False Negative (FN)
Actual Negative	False Positive (FP)	True Negative (TN)

Table 3.2 show the confusion matrix where: True Positives are the number of correctly predicted positive instances. False Positives is the number of incorrectly predicted positive instances, True Negative is the number of correctly predicted negative cases while False Negative is the number of incorrectly predicted negative instances. Accuracy and precision of the methods can be evaluated by using confusion matrix. Accuracy is the percentage of accurate predictions the model makes is measured and the formula is (TP+TN)/(TP+TN+FP+FN). Meanwhile, precision is the ratio of accurate positive predictions to all positive predictions made by the model is measured and the formula is TP/(TP+FP).

3.5.2 Biological Context Verification

The goal of this validation procedure is to make sure whether there is present study or other proof linking the identified gene to the targeted potential biomarkers of EC. Author wished to verify the potential significance of the identified genes and increase the confidence in our findings by undertaking a thorough search. The

biological context validation stage ensures that the genes discovered are not simply based on their existence in the biclusters but are also supported by scientific data in the literature. With a more solid foundation for further evaluation and interpretation, the outcomes are more reliable and legitimate because of this thorough methodology.

3.5.3 Sum of Square Method

The sum of square is a method to calculate the dispersion of data points around the mean (Nainggolan et al, 2019). The formula of the sum of square is as below.

$$SSE = \sum_{i=0}^{n} (X_i - \bar{X})^2 \tag{3.1}$$

Where:

SSE: sum of squared error

$\sum_{i=0}^{n}()$: summation of the data.

X_i: means values the *i*th data.

$\bar{X}$: means values for all data.

According to the equation above, the data will be used to calculate the mean value of each row and obtain the mean value for all the data. By subtracting the rows' mean value with the mean value for all the data, getting the square of the differences and summing them together, the SSE value for the data will be able to obtain.

Larger SSE values typically imply greater dispersion or variability inside the cluster, which suggests that the data points are more dispersed and possibly not adequately clustered. In contrast, a low SSE score suggests that there is little dispersion or variability inside the cluster which indicates that the data points are more closely grouped in their respective clusters.

3.6 Hardware and Software Requirements

This project requires Microsoft Visual Studio and R Studio. Python code can be developed using Microsoft Visual Studio. On the other hand, R Studio is an integrated development environment for R programming. Microsoft Excel needed to be used for analyzing data.

Specific hardware needs must be considered for this study to ensure efficient analysis and minimized time complexity. The minimum hardware requirements for this study are RAM 4GB, Intel Core i5 Processor and Windows 10 operating system.

3.7 Chapter Summary

In a nutshell, this chapter explained the research framework as well as the activities needed to be done in each phase. The author will consider all the phases to achieve the goals of this research. The datasets used for this study have been illustrated and explained. The measurement of the effectiveness of the algorithms to identify the potential biomarkers had been shown in this chapter as well as the hardware and software requirements needed for efficiency analyzation. Next chapter will discuss the development of the proposed biclustering methods.

CHAPTER 4

RESEARCH DESIGN AND IMPLEMENTATION

4.1 Introduction

In this chapter, a step-by-step procedure had been laid out for identifying possible biomarkers for endometrial cancer (EC), starting with dataset preparation, and ending with validation. Finding genes with a strong association to EC and the potential to act as biomarkers for the condition is the aim.

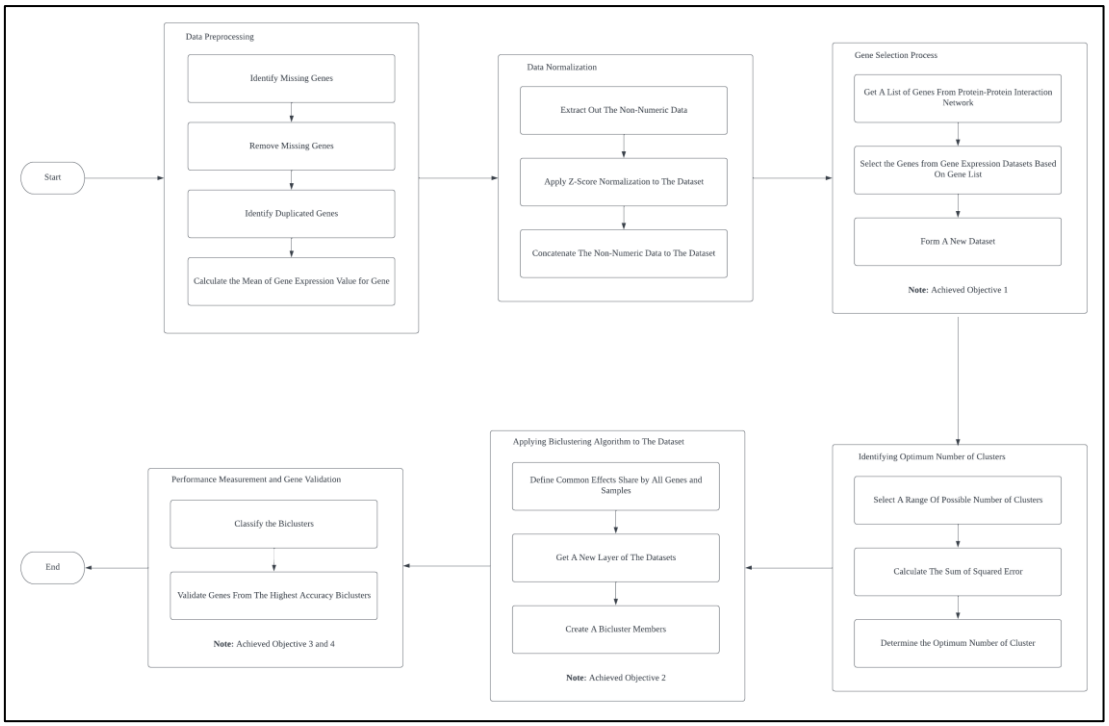


Figure 4.1: Development Process

4.2 Data Preparation

Data preparation is an essential stage in data analysis because data must be transformed into a format that can be used for analysis, modelling, and interpretation. In this research, there are two datasets that will be used for further analysis. Both datasets were obtained from GEO and STRING respectively. The GEO dataset contains 22278 rows of genes, and 34 columns of samples. For STRING, there are 3506 human genes showing the relationship between each other.

4.2.1 Data Pre-processing

Data preprocessing is an important step in ensuring the input data is clean and formatted before analyzing. The missing genes in the datasets had been removed and eliminated to improve computational efficiency. Besides that, the genes which occurred more than one will then be calculated to obtain the mean values. The dimension of the dataset which had been removed the missing genes and obtained the mean values of the duplicated genes became 13514 genes x 34 samples.

4.2.2 Data Normalization

A gene's Z-score can be computed by comparing its expression level in one sample to its expression level in all samples (Pluto Bioinformatics, 2022). Hence, Z-score normalization had been applied to the gene expression datasets. The values after normalization normally fall between 0 and 1. Normalizing the data ensures that each variable contributes equally to the analysis. By arranging all variables on a comparable scale, linkages and patterns in the data become more visible and interpretable.

$$Z = \frac{x - \mu}{\delta} \tag{4.1}$$

Where:

Z: standard score

x: observed values.

μ: *means of the sample*

δ: standard deviation of the sample

	Gene Symbol	GSM509788_E1520N.CEL	GSM509789_E1521N.CEL	GSM509790_E1532N.CEL	GSM509791_E1535N.CEL	GSM509792_E1542N.CEL
0	DDR1 /// MIR4640	1.679350	1.534117	1.641309	1.592089	1.415333
1	RFC2	0.177316	0.267164	0.190596	0.107687	-0.128427
2	HSPA6	-0.398463	-0.486782	-0.428942	-0.281714	-0.339854
3	PAX8	-0.394187	-0.381440	-0.386848	-0.385943	-0.389198
4	GUCA1A	-1.193654	-1.246400	-1.166557	-1.052665	-1.185264
...	...	...	...	...	...	...
13510	FAM86B1 /// FAM86B2 /// FAM86C1 /// FAM86DP //...	0.121828	0.203201	0.061114	-0.104431	0.080674
13511	SNHG17	-0.002660	0.255355	0.026558	-0.018404	0.012422
13512	HNRNPUL2 /// HNRNPUL2-BSCL2	-1.145683	-1.294155	-1.079146	-1.108209	-1.177337
13513	LOC100505915	-0.719130	-0.646220	-0.787132	-0.426511	-0.713784
13514	NPEPL1	0.430177	0.390383	0.476316	0.454189	0.437905

13515 rows × 34 columns

Figure 4.2: Gene Expression Datasets After Normalization

4.2.3 Gene Selection Process

The human genes in the PPI network are then to be retrieved and act as secondary genes data. The genes in the gene expression data act as primary genes data. Then, after we retrieved, human genes are then used to select the genes in the gene expression data. Hence, the new dataset contained only the genes which occurred in the gene expression data and PPI network. After gene selection, the dimensions of the datasets will be 2735 genes x 34 samples.

Further explanation of gene selection process can be referenced on the Figure 4.3. A gene list had been generated from PPI data. Gene 1, Gene 2 and Gene 4 from gene expression dataset had been selected and form an input data. This is because Gene

1, Gene 2 and Gene 4 were found in the gene list while Gene 3 was not found in the gene list and been eliminated to form the input data.

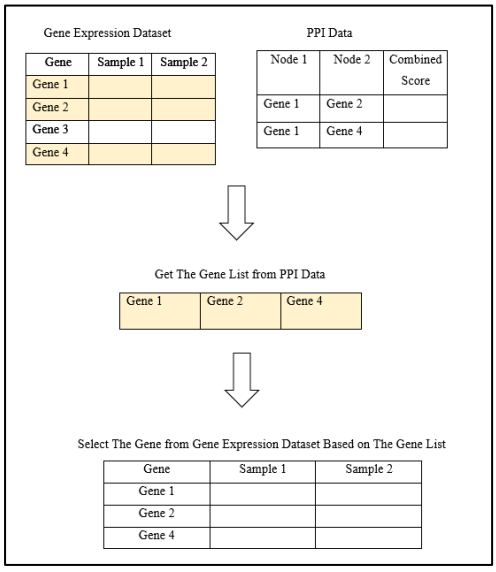


Figure 4.3: The Workflow of Gene Selection Process

	Gene Symbol	GSM509788_E1520N.CEL	GSM509789_E1521N.CEL	GSM509790_E1532N.CEL	GSM509791_E1535N.CEL	GSM509792_E1542N.CEL
0	HSPA6	-0.398463	-0.486782	-0.428942	-0.281714	-0.339854
1	GUCA1A	-1.193654	-1.246400	-1.166557	-1.052665	-1.185264
2	CCL5	0.548768	0.758375	0.071602	0.632368	0.374798
3	MMP14	0.162577	0.128637	0.179595	0.270355	0.163781
4	TRADD	0.213093	0.255694	0.352167	0.296220	0.206041
...	...	...	...	...	...	...
2730	TRDN	-1.613432	-1.481663	-1.592278	-1.526531	-1.622887
2731	SCAF4	-0.442155	-0.013311	-0.186452	0.275966	-0.281995
2732	LAMA1	-0.905482	-0.884916	-0.974489	-0.918797	-0.987055
2733	FBXO31	0.138107	-0.010299	0.003653	-0.133142	0.011334
2734	SLC44A1	0.002750	-0.251866	-0.118740	0.252419	-0.220769

2735 rows × 34 columns

Figure 4.4: The Input Data

4.3 Identify the Optimum Number of Clusters

The elbow method is a technique which is used to find the optimal number of clusters. The concept is finding the elbow point of the sum of squared error and the number of clusters. Sum of squared error is the sum of squared distance for each data point.

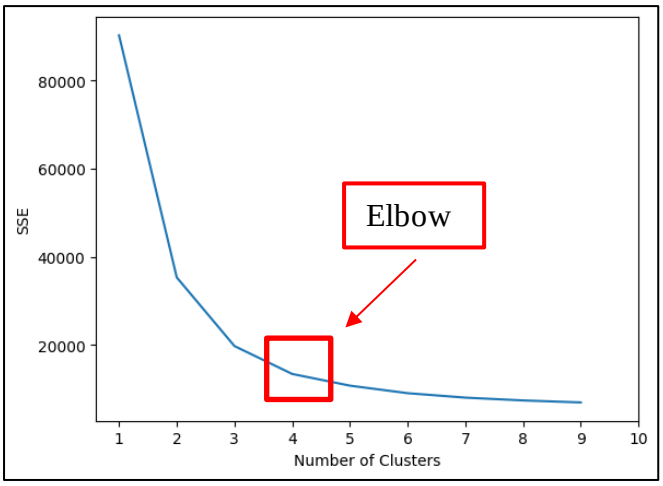


Figure 4.5: Elbow Method

4.4 Applying Biclustering Algorithm

Figure 4.5 shows the general flows of the Plaid Biclustering Method. The explanation of each step will then be further discussed.

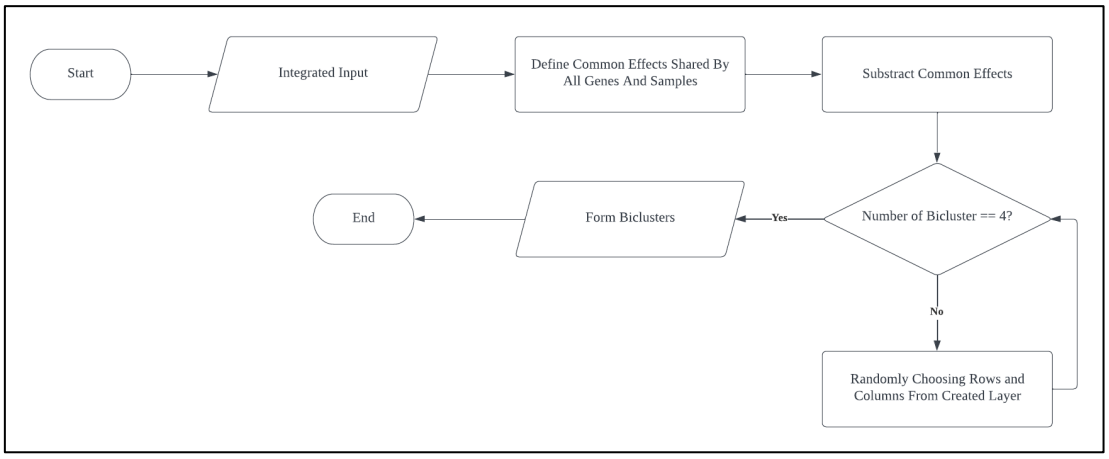


Figure 4.6: Basic Architecture of Plaid Biclustering

4.4.1 Create Layer from Residuals for Pattern Capture

There is a background layer in the Plaid bicluster model. Background layers in Plaid models indicate common effects shared by all genes and samples; by adding new layers, particular effects can be separated from background layers to show biclusters

that are specific to a condition or treatment. In this step, the mean, row effects and column effects of the input data will be calculated. This method captures both the overall average behavior of the row and column divergent by computing row and column effects.

4.4.2 Substract Background Layer/Common Effects

By subtracting the background layer from the residuals, the algorithm effectively removes the common impact represented by the background layer from the residuals. This procedure updates the residuals to concentrate on any remaining precise changes that the background layer is unable to account for.

4.4.3 Formed A Collection of Biclusters

Fitting new layers in the biclustering method is critical. This technique involves selecting relevant rows and columns that capture particular changes in the gene expression data as you iterate through stages. As a result, particular rows and columns are chosen to create double cluster members. As a result of this, the data may overlap in several biclusters. When examining biomarkers in EC, obtaining overlapping double clusters is especially beneficial since it increases the chances of discovering similar gene expression patterns.

4.5 Performance Measurements and Gene Validation

4.5.1 Classification of the Biclusters

After we retrieve the biclusters from the Plaid Biclustering Algorithm, the biclusters members will then be separated into three categories. A suitable validation

and testing procedure, such as cross-validation or independent testing, can help assess classification performance and identify the most effective approach. The classification method that will be used is SVM. Then, the performance of each classification method for each category will be evaluated by confusion matrix.

Table 4.1: Categories of Bicluster Members

Category A	Biclusters With Largest Size
Category B	All Biclusters
Category C	Combination of the Biclusters

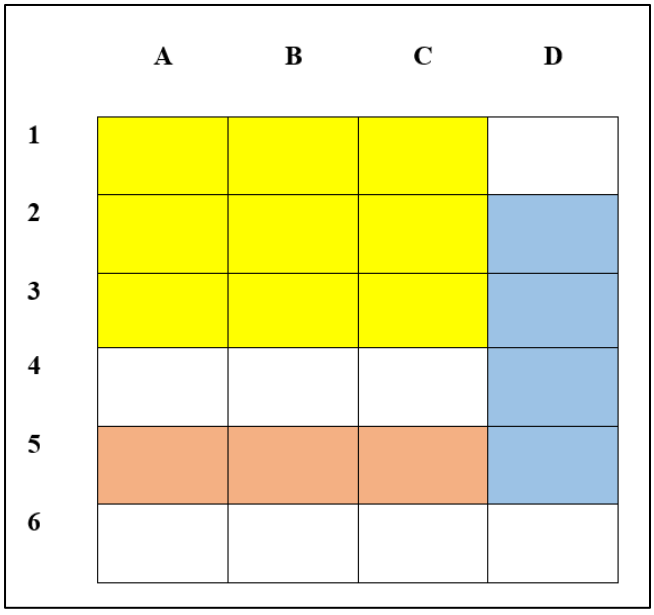


Figure 4.7: Sample of Biclusters for Classification

Table 4.1 will be further explained using Figure 4.6 as an example. Three biclusters are shown in Figure 4.6, each with a unique set of dimensions. Blue biclusters are 4x1, yellow biclusters are 3x3 and light brown biclusters are 1x3. The biggest yellow bicluster, which has the dimension of 3x3 is refer as Category A. Then, Category B is made up of the yellow, light brown, and blue biclusters. The biclusters are further arranged for Category C as illustrated in Table 4.2.

Table 4.2: Combination of the Biclusters

Combination 1	Yellow Bicluster, Light Brown Bicluster
Combination 2	Yellow Bicluster, Blue Bicluster
Combination 3	Light Brown Bicluster, Blue Bicluster

4.5.2 Verify the Selected Potential Biomarkers

The bicluster which achieved the better results on classification will then be identified as the potential biomarkers for EC. However, using only statistical results to point out the biomarkers for EC is not enough, hence the genes in the bicluster will then be further verified with the biological knowledgebases.

4.6 Chapter Summary

In conclusion, there are numerous critical phases involved in the process of finding possible EC biomarkers. The Plaid biclustering algorithm was used to extract four biclusters, each of which contained a number of members. The final objective of the research is to achieve a classification accuracy of more than 85% when various classification models are applied to different categories. Additionally, the chosen potential biomarkers must show a strong correlation with EC, demonstrating their applicability in the context of the disease. This study seeks to understand and identify useful biomarkers for EC detection and diagnosis using the criteria and procedures outlined above.

CHAPTER 5

CONCLUSION AND RECOMMENDATIONS

5.1 Research Outcomes

The input data is successfully extracted from the gene expression dataset and PPI dataset. By using Elbow Method, the optimum number of biclusters for Plaid model is four. The biclustering method successfully identified four different groups of biclusters based on the input data. Then, classification model that had been apply to the bicluster was predicted with more than 85% accuracy. In order to confirm the found biomarkers' relationship with disease, established biological knowledgebases were used to cross-validate a group of genes that have the highest accuracy. The potential biomarkers that have been found are ARPC2, APPL1, FTL, and PLAU.

5.2 Achievements

Achievements for this research are:

(a) Data on gene expression and protein-protein interactions (PPI) can be effectively combined to produce input datasets for analysis

(b) Number of biclusters was determined optimally by using the Elbow Method

5.3 Future Works

Future works in PSM II are:

(a) Implementation of the Biclustering Algorithm to identify four different sets of biclusters from an input dataset.

(b) Classification models such as SVM, Random Forest, Decision Tree and kNN have been applied to the biclusters. The found biomarkers have the potential to be used in the early identification of esophageal cancer, as shown by the classification model's accuracy.

(c) Validation the correlation and association of found biomarkers with esophageal cancer by cross-validating them with reputable biological knowledge bases like NCBI and UniProt.

REFERENCES

Abd-Elnaby, M., Alfonse, M. and Roushdy, M. (2021) 'Classification of breast cancer using microarray gene expression data: A survey', *Journal of biomedical informatics*, 117, p.103764.

Almugren, N. and Alshamlan, H.M. (2019) 'New bio-marker gene discovery algorithms for cancer gene expression profile', *IEEE Access*, 7, pp.136907-136913.

Arnold, M., Soerjomataram, I., Ferlay, J. and Forman, D. (2015) 'Global incidence of oesophageal cancer by histological subtype in 2012', *Gut*, 64(3), pp.381-387.

Athanasios, A., Charalampos, V. and Vasileios, T. (2017) 'Protein-protein interaction (PPI) network: recent advances in drug discovery', *Current drug metabolism*, 18(1), pp.5-10.

Ayyad, S.M., Saleh, A.I. and Labib, L.M. (2019) 'Gene expression cancer classification using modified K-Nearest Neighbors technique', *Biosystems*, 176, pp.41-51.

Branders, V., Schaus, P. and Dupont, P. (2019) 'Identifying gene-specific subgroups: an alternative to biclustering', *BMC bioinformatics*, 20(1), pp.1-13.

Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A. (2018) 'Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries.' *CA: a cancer journal for clinicians*, 68(6), pp.394-424.

Bustamam, A., Siswantining, T., Kaloka, T.P. and Swasti, O. (2020) 'Application of bimax, pols, and lcm-mbc to find bicluster on interactions protein between hiv-1 and human', *Austrian Journal of Statistics*, 49(3), pp.1-18.

Cabri, W., Cantelmi, P., Corbisiero, D., Fantoni, T., Ferrazzano, L., Martelli, G., Mattellone, A. and Tolomelli, A. (2021) 'Therapeutic peptides targeting PPI in clinical development: Overview, mechanism of action and perspectives', *Frontiers in Molecular Biosciences*, 8, p.697586.

Castanho, E.N., Aidos, H. and Madeira, S.C. (2022) 'Biclustering fMRI time series: a comparative study', *BMC bioinformatics*, 23(1), pp.1-30.

Charbuty, B. and Abdulazeez, A. (2021) 'Classification based on decision tree algorithm for machine learning.' *Journal of Applied Science and Technology Trends*, 2(01), pp.20-28.

Cui, Y., Zhang, R., Gao, H., Lu, Y., Liu, Y. and Gao, G. (2020) 'A novel biclustering of gene expression data based on hybrid BAFS-BSA algorithm', *Multimedia Tools and Applications*, 79, pp.14811-14824.

Czajkowski, M. and Kretowski, M. (2019) 'Decision tree underfitting in mining of gene expression data. An evolutionary multi-test tree approach', *Expert Systems with Applications*, 137, pp.392-404.

Di Iorio, J., Chiaromonte, F. and Cremona, M.A. (2020) 'On the bias of H-scores for comparing biclusters, and how to correct it', *Bioinformatics*, 36(9), pp.2955-2957.

Do, K.A., Müller, P. and Tang, F. (2005) 'A Bayesian mixture model for differential gene expression', *Journal of the Royal Statistical Society Series C: Applied Statistics*, 54(3), pp.627-644.

Eren, K. (2012) *Application of Biclustering Algorithm To Biological Data*. Master Thesis, The Ohio State University, United States.

Eren, K., Deveci, M., Küçüktunç, O. and Çatalyürek, Ü.V. (2013) 'A comparative analysis of biclustering algorithms for gene expression data', *Briefings in bioinformatics*, 14(3), pp.279-292.

Freitas, A., Afreixo, V., Pinheiro, M., Oliveira, J.L., Moura, G. and Santos, M. (2011) 'Improving the performance of the iterative signature algorithm for the identification of relevant patterns', *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 4(1), pp.71-83.

Hayasaka S. (2022). *How Many Cluster? Methods for choosing the right number of clusters.* Towards Data Science. Available at: https://towardsdatascience.com/how-many-clusters-6b3f220f0ef5#:~:text=The%20silhouette%20coefficient%20may%20provide,peak%20as%20the%20optimum%20K. (Accessed on 3 July 2023)

Henriques, R. and Madeira, S.C. (2015) 'BicNET: efficient biclustering of biological networks to unravel non-trivial modules', *In Algorithms in Bioinformatics: 15th International Workshop, WABI 2015, Atlanta, GA, USA, September 10-12, 2015, Proceedings 15*, pp. 1-15.

Karim, M.B., Kanaya, S. and Altaf-Ul-Amin, M. (2019) 'Implementation of BiClusO and its comparison with other biclustering algorithms', *Applied Network Science*, 4(1), pp.1-15.

Karimizadeh, E., Sharifi-Zarchi, A., Nikaein, H., Salehi, S., Salamatian, B., Elmi, N., Gharibdoost, F. and Mahmoudi, M. (2019) 'Analysis of gene expression profiles and protein-protein interaction networks in multiple tissues of systemic sclerosis', *BMC medical genomics*, *12*, pp.1-12.

Kocatürk, A., Altunkaynak, B. and Homaida, A. (2019) 'Comparing Biclustering Algorithms Using Data Envelopment Analysis to Choose the Best Parameters', In *2019 International Artificial Intelligence and Data Processing Symposium (IDAP) IEEE*. pp. 1-14.

Komorowski, M., Green, A., Tatham, K.C., Seymour, C. and Antcliffe, D. (2022) 'Sepsis biomarkers and diagnostic tools with a focus on machine learning', *EBioMedicine*, p.104394.

Kumar A. (2021). *Elbow Method vs Silhouette Score – Which is better?* Data Analystics. Available at: https://vitalflux.com/elbow-method-silhouette-score-which-better/#:~:text=The%20calculation%20simplicity%20of%20elbow,k%20that%20is%20the%20best. (Accessed on: 3 July 2023)

Lagergren, J., Smyth, E., Cunningham, D. and Lagergren, P. (2017) 'Oesophageal cancer', *The Lancet*, 390(10110), pp.2383-2396.

Li, G. (2020) UniBic: An Elementary Method Revolutionizing Biclustering. Hyderabab, India: Vide Leaf. 2020.

Liu, F., Yang, Y., Xu, X.S. and Yuan, M. (2022) 'Mutually exclusive spectral biclustering and its applications in cancer subtyping', *bioRxiv*, pp.1-29.

Liu, X., Li, D., Liu, J., Su, Z. and Li, G. (2020) 'RecBic: a fast and accurate algorithm recognizing trend-preserving biclusters', *Bioinformatics*, 36(20), pp.5054-5060.

Lu, H., Zhou, Q., He, J., Jiang, Z., Peng, C., Tong, R. and Shi, J. (2020) 'Recent advances in the development of protein–protein interactions modulators: mechanisms and clinical trials', *Signal transduction and targeted therapy*, 5(1), p.213.

Luque, A., Carrasco, A., Martín, A. and de Las Heras, A. (2019) 'The impact of class imbalance in classification performance metrics based on the binary confusion matrix', *Pattern Recognition*, 91, pp.216-231.

Maind, A. and Raut, S. (2019) 'COSCEB: Comprehensive search for column-coherent evolution biclusters and its application to hub gene identification', *Journal of biosciences*, 44, pp.1-16.

Meeds, E. and Roweis, S. (2007) 'Nonparametric bayesian biclustering', *Technical report UTML TR 2007-001*, 2007 (June), pp 1-12.

Moteghaed, N.Y., Maghooli, K. and Garshasbi, M. (2018) 'Improving classification of Cancer and mining biomarkers from gene expression profiles using hybrid optimization algorithms and fuzzy support vector machine', *Journal of medical signals and sensors*, 8(1), p.1

Nainggolan, R., Perangin-angin, R., Simarmata, E. and Tarigan, A.F. (2019) 'Improved the performance of the K-means cluster using the sum of squared error (SSE) optimized by using the Elbow method', In *Journal of Physics: Conference Series* , p. 012015.

Napier, K.J., Scheerer, M. and Misra, S. (2014) 'Esophageal cancer: A Review of epidemiology, pathogenesis, staging workup and treatment modalities', *World journal of gastrointestinal oncology*, 6(5), p.112.

National Cancer Institution. (no date). *NCI Dictionary of Cancer Terms.* Available at: https://www.cancer.gov/publications/dictionaries/cancer-terms/def/biomarker (Accessed: 8 April 2023).

National Human Genome Research Institute. (2023). *Gene Expression*. Available at: https://www.genome.gov/genetics-glossary/Gene-Expression#:~:text=Gene%20expression%20is%20the%20process,molecules%20that%20serve%20other%20functions. (Accessed on 12 May 2023).

Otchere, D.A., Ganat, T.O.A., Gholami, R. and Ridha, S. (2021) 'Application of supervised machine learning paradigms in the prediction of petroleum reservoir properties: Comparative analysis of ANN and SVM models', *Journal of Petroleum Science and Engineering*, 200, p.108182.

Patowary, P. and Bhattacharyya, D.K. (2021) 'PD_BiBIM: Biclustering-based biomarker identification in ESCC microarray data', *Journal of Biosciences*, 46(3), p.56.

Pinto, H., Gates, I. and Wang, X. (2020) 'Bayesian biclustering by dynamics: Algorithm testing, comparison against random agglomeration, and calculation of application specific prior information', *MethodsX*, 7, p.100897.

Pluto Bioinformatics. (2022). *Understanding the Z-scores in RNA seq Analysis.* Available at https://pluto.bio/blog/overview-of-z-scores-in-rna-seq-experiments (Accessed on 17 May 2023).

Rai, V., Abdo, J. and Agrawal, D.K. (2023) 'Biomarkers for Early Detection, Prognosis, and Therapeutics of Esophageal Cancers', *International Journal of Molecular Sciences*, 24(4), p.3316.

Rao, V.S., Srinivas, K., Sujini, G.N. and Kumar, G.N. (2014) 'Protein-protein interaction detection: methods and analysis', *International journal of proteomics*, 2014, p.147168.

Rashidi, H.H., Khan, I.H., Dang, L.T., Albahra, S., Ratan, U., Chadderwala, N., To, W., Srinivas, P., Wajda, J. and Tran, N.K. (2022) 'Prediction of tuberculosis using an automated machine learning platform for models trained on synthetic data', *Journal of pathology informatics*, 13, p.100172.

Ray, S. (2019) 'A quick review of machine learning algorithms', In 2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon), pp. 35-39.

Renc, P., Orzechowski, P., Byrski, A., Wăs, J. and Moore, J.H. (2021) 'EBIC. JL: an efficient implementation of evolutionary biclustering algorithm in Julia', *In Proceedings of the Genetic and Evolutionary Computation Conference Companion*, pp. 1540-1548.

Shaharudin, S.M., Ismail, S., Nor, S.M.C.M. and Ahmad, N. (2019) 'An efficient method to improve the clustering performance using hybrid robust principal component analysis-spectral biclustering in rainfall patterns identification', *IAES International Journal of Artificial Intelligence*, 8(3), p.237.

Siswantining, T., Aminanto, A.E., Sarwinda, D. and Swasti, O. (2021) 'Biclustering Analysis Using Plaid Model on Gene Expression Data of Colon Cancer', *Austrian Journal of Statistics*, 50(5), pp.101-114.

Steardo Jr, L., Carbone, E.A., De Filippis, R., Pisanu, C., Segura-Garcia, C., Squassina, A., De Fazio, P. and Steardo, L. (2020) 'Application of support vector machine on fMRI data as biomarkers in schizophrenia diagnosis: a systematic review', *Frontiers in Psychiatry*, 11, p.588.

Supper, J., Strauch, M., Wanke, D., Harter, K. and Zell, A. (2007) 'EDISA: extracting biclusters from multiple time-series of gene expression profiles', *BMC bioinformatics*, 8, pp.1-14.

Sutheeworapong, S., Ota, M., Ohta, H. and Kinoshita, K. (2012) 'A novel biclustering approach with iterative optimization to analyze gene expression data', *Advances and Applications in Bioinformatics and Chemistry*, pp.23-59.

Tanay, A., Sharan, R. and Shamir, R. (2005) 'Biclustering algorithms: A survey', *Handbook of computational molecular biology*, 9(1-20), pp.122-124.

Uddin, S., Khan, A., Hossain, M.E. and Moni, M.A. (2019) 'Comparing different supervised machine learning algorithms for disease prediction', *BMC medical informatics and decision making*, 19(1), pp.1-16.

Voggenreiter, O., Bleuler, S. and Gruissem, W. (2012) 'Exact biclustering algorithm for the analysis of large gene expression data sets', *BMC bioinformatics*, 13(Suppl 18), p.10.

Wang, M., Smith, J.S. and Wei, W.Q. (2018) 'Tissue protein biomarker candidates to predict progression of esophageal squamous cell carcinoma and precancerous lesions', *Annals of the New York Academy of Sciences*, 1434(1), pp.59-69.

Wang, Y., Zhao, Y., Therneau, T.M., Atkinson, E.J., Tafti, A.P., Zhang, N., Amin, S., Limper, A.H., Khosla, S. and Liu, H. (2020) 'Unsupervised machine learning for the discovery of latent disease clusters and patient subgroups using electronic health records', *Journal of biomedical informatics*, 102, p.103364.

World Cancer Research Fund International. (no date). *Oesophageal cancer statistics*. Available at: https://www.wcrf.org/cancer-trends/oesophageal-cancer-statistics/ (Accessed: 12 May 2023).

Xie, J., Ma, A., Fennell, A., Ma, Q. and Zhao, J. (2019) 'It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data', *Briefings in bioinformatics*, 20(4), pp.1450-1465.

Xie, J., Ma, A., Zhang, Y., Liu, B., Cao, S., Wang, C., Xu, J., Zhang, C. and Ma, Q. (2020) 'QUBIC2: a novel and robust biclustering algorithm for analyses and interpretation of large-scale RNA-Seq data', *Bioinformatics*, 36(4), pp.1143-1149.

Xie, Y., Meng, W.Y., Li, R.Z., Wang, Y.W., Qian, X., Chan, C., Yu, Z.F., Fan, X.X., Pan, H.D., Xie, C. and Wu, Q.B. (2021) 'Early lung cancer diagnostic

biomarker discovery by machine learning methods', *Translational oncology*, 14(1), p.100907.

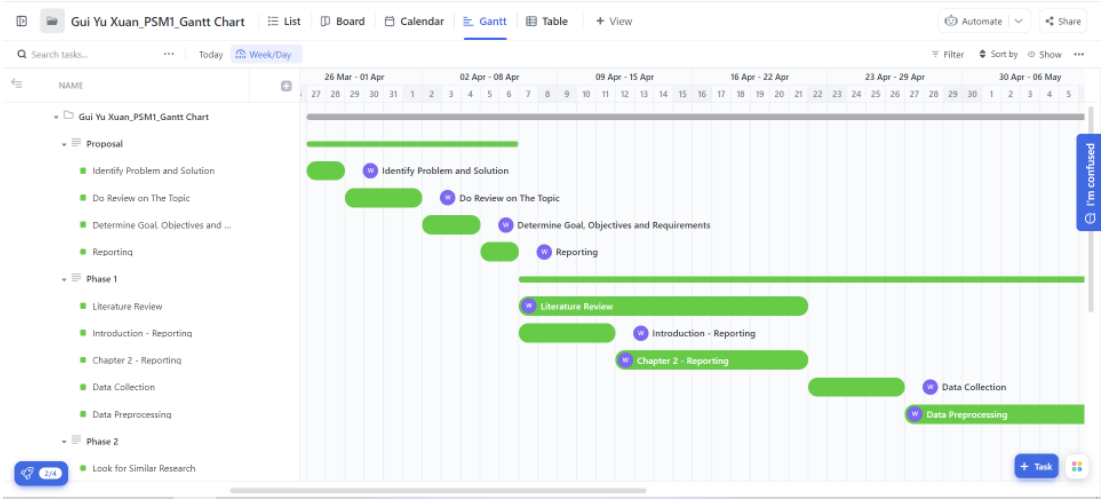
Yang, J., Liu, X., Cao, S., Dong, X., Rao, S. and Cai, K. (2020) 'Understanding esophageal cancer: the challenges and opportunities for the next decade', *Frontiers in oncology*, 10, p.1727.

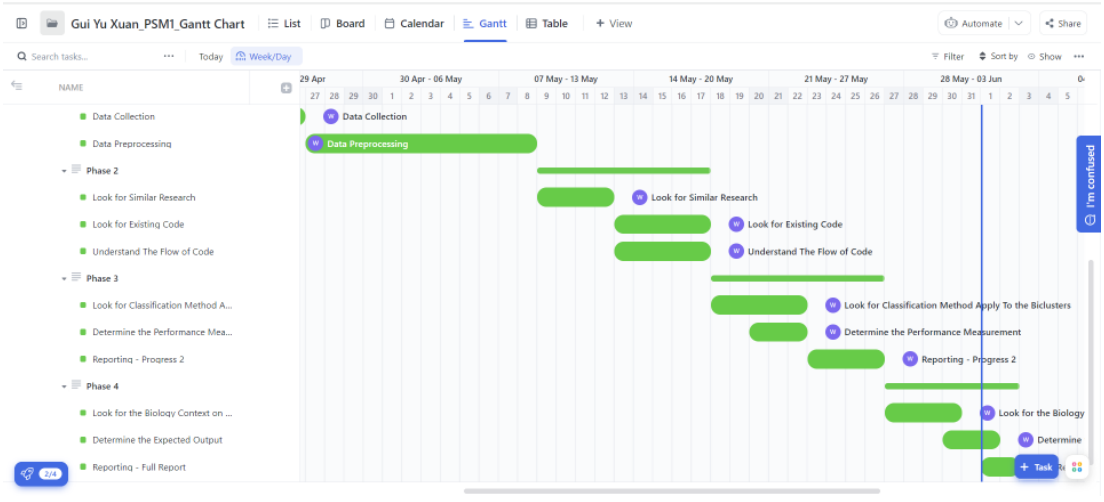
Yang, J., Wang, H., Wang, W. and Yu, P. (2003) Enhanced biclustering on expression data, In *Third IEEE Symposium on Bioinformatics and Bioengineering, 2003. Proceedings*, pp. 321-327.

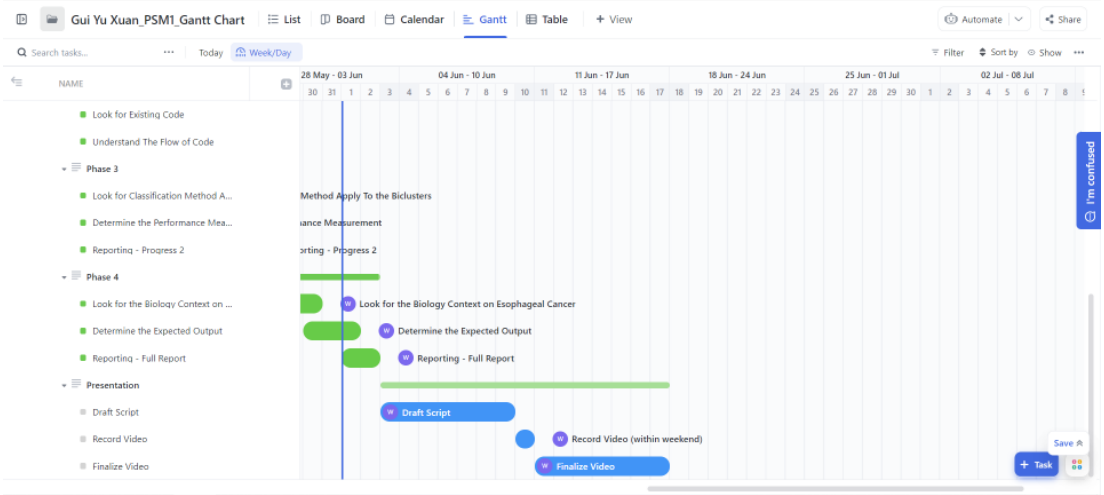
Yousef, M., Kumar, A. and Bakir-Gungor, B. (2020) 'Application of biological domain knowledge based feature selection on gene expression data', *Entropy*, 23(1), p.2.

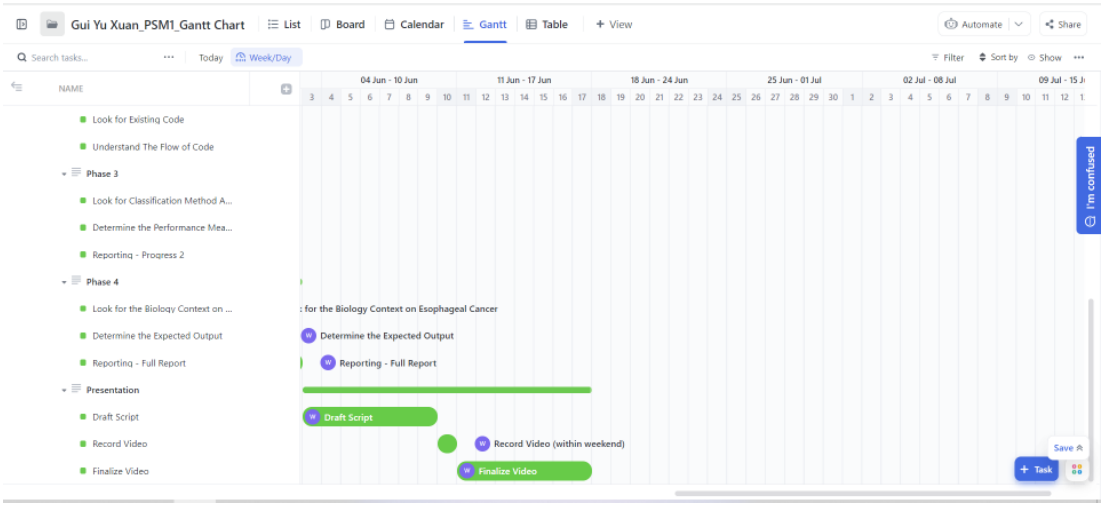
Appendix A A Gantt Chart for PSM 1

https://sharing.clickup.com/25555781/g/h/rbwu5-928/f2c01cb149b1de5








Gui Yu Xuan_PSM1_Gantt Chart
List
Board
Calendar
Gantt
Table
View
Automate
Share
Search tasks...
Today
Week/Day
Filter
Sort by
Show
NAME
04 Jun - 10 Jun
11 Jun - 17 Jun
18 Jun - 24 Jun
25 Jun - 01 Jul
02 Jul - 08 Jul
Look for Existing Code
Understand The Flow of Code
Phase 3
Look for Classification Method A...
Determine the Performance Mea...
Reporting - Progress 2
Phase 4
Look for the Biology Context on ...
Determine the Expected Output
Reporting - Full Report
Presentation
Draft Script
Record Video
Finalize Video
: for the Biology Context on Esophageal Cancer
Determine the Expected Output
Reporting - Full Report
Draft Script
Record Video (within weekend)
Finalize Video
I'm confused
Save
Task