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 : 20232024 - 02

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

A20EC0039

MATRIC NUMBER

Date: 1 JULY 2024

SIGNATURE OF SUPERVISOR

DR. CHAN WENG HOWE

NAME OF SUPERVISOR

Date: 1 JULY 2024

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 : 1 JULY 2024

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

JULY 2024

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 : 1 JULY 2024

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. Various biclustering methods had been proposed to identify the potential biomarkers for certain diseases. However, most research were done based on the synthetic 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 was to implement a biclustering method to identify the potential biomarkers of esophageal cancer from gene expression data and protein-protein interaction data. In this research, the gene expression dataset and protein-protein interaction datasets were used in the gene selection process and applied in the biclustering method. Elbow method had been used to determine the optimum number of biclusters. Four bicusters were obtained in this study, each bicluster will then be observed and the genes of the biclusters were used to filter the gene expression dataset. The biclustering method used in this research was Plaid model, which selected the rows and columns exhibiting the similar pattern from the dataset to form biclusters. The results obtained from the biclustering algorithm indicated that the biclusters formed consisted only of cancerous cases, making them unsuitable for implementation with the Support Vector Machine classifier. Thus, the genes were examined and formed different type of Gene Expression Dataset for comparison. Subsequently, different Gene Expression Dataset were classified by the Support Vector Machine. Two datasets were formed, one involving genes in all biclusters and another involving genes that occurred in more than one biclusters. The Support Vector Machine was implemented on these datasets along with the original gene expression dataset with accuracy of 96.43%, 95% and 96.43% respectively. The dataset involving genes that occurred in more than one bicluster was validated with biological knowledgebases. The potential biomarkers for the esophageal cancer found in the experiment are EPHB4, LAMB3 and HOXD11. To conclude, the potential biomarkers for esophageal cancer found in this research have the potential to improve the early detection and diagnosis for esophageal cancer and improve in the available treatments.

ABSTRAK

Biclustering ialah cara analisis data untuk mengumpulkan kluster berdasarkan ciri-ciri tertentu. Walaupun terdapat pelbagai kaedah biclustering, namun kebanyakan penyelidikan dijalankan menggunakan data sintetik yang mungkin menghasilkan keputusan yang positif palsu dan terlalu sesuai dengan data. Oleh itu, kekurangan data yang relevan secara biologi dalam analisis biclustering menyebabkan ketepatan yang rendah dalam mengenal pasti kluster gen yang relavan dan mengurangkan ketepatan pengesanan biomarker. Tujuan kajian ini adalah untuk melaksanakan kaedah biclustering untuk mengenal pasti biomarker yang berpotensi untuk kanser esofagus daripada data ekspresi gen dan interaksi protein-protein. Dalam kajian ini, data ekspresi gen dan data interaksi protein-protein digunakan dalam proses pemilihan gen dan diaplikasikan dalam kaedah biclustering. Kaedah elbow telah digunakan untuk menentukan bilangan kluster yang optimum. Empat bicluster diperoleh dalam kajian ini, dan gen-gen dalam bicluster digunakan untuk menapis dataset ekspresi gen. Kaedah biclustering yang digunakan ialah model Plaid, yang memilih baris dan lajur yang menunjukkan corak yang serupa dari dataset untuk membentuk bicluster. Hasil yang diperoleh menunjukkan bahawa bicluster yang terbentuk terdiri daripada kes-kes kanser sahaja, menjadikannya tidak sesuai dengan pengelasan Mesin Sokong Vektor. Oleh itu, gen-gen diperiksa dan pelbagai data ekspresi gen diklasifikasikan oleh Mesin Sokongan Vektor. Dua dataset dibentuk, satu melibatkan gen dalam semua bicluster dan satu melibatkan gen yang terdapat dalam lebih daripada satu bicluster. Mesin Sokongan Vektor diterapkan pada data tersebut bersama dengan data ekspresi gen asal dengan ketepatkan masing-masing sebanyak 96.43%, 95% dan 96.43%. Dataset yang melibatkan gen yang terdapat dalam lebih daripada satu bicluster disahkan dengan pangkalan pengetahuan biologi. Biomarker berpotensi untuk kanser esofagus yang ditemui dalam eksperimen adalah EPHB4, LAMB3 dan HOXD11. Secara kesimpulannya, biomarker berpotensi untuk kanser esofagus yang ditemui dalam penyelidikan ini mempunyai potensi untuk meningkatkan pengesanan awal dan diagnosis untuk kanser esofagus serta meningkatkan 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

$\sum$	-	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 specific to EC cancer.

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. For example, biclustering grouped genes and samples based on their patterns by finding the co-expressed subset of data.

Even the performance of classification can be improved by the biclustering algorithm, but the biclustering algorithm still had limitations to run the experiment. 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). However, using only synthetic data to find biomarkers can produce false-positive results and overfit the data (Rashidi et al, 2022). Synthetic data is made by combining real world information to create a dataset that resembles actual data but does not reveal any personal information (Rashidi et al, 2022). As synthetic data do not capture full patterns present in real world data, thus the result obtained may not be accurately and caused overfitting. 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). Hence, 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. 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 implement 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 derive input data from esophageal cancer gene expression and protein-protein interaction data.

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

(c) To evaluate the selected potential biomarkers using Support Vector Machine through ten-fold cross validation and confusion matrix.

(d) To verify the identified potential biomarkers with biological knowledgebases such as NCBI.

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.

(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 an effective biclustring method, the accuracy and reliability of biomarker identification would be improved. 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 were 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 were explained in further. The popular biclustering algorithms were discussed and the advantageous and drawbacks of each algorithm were outlined.

2.2 Esophageal Cancer (EC)

EC has been reported as 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 displayed 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).

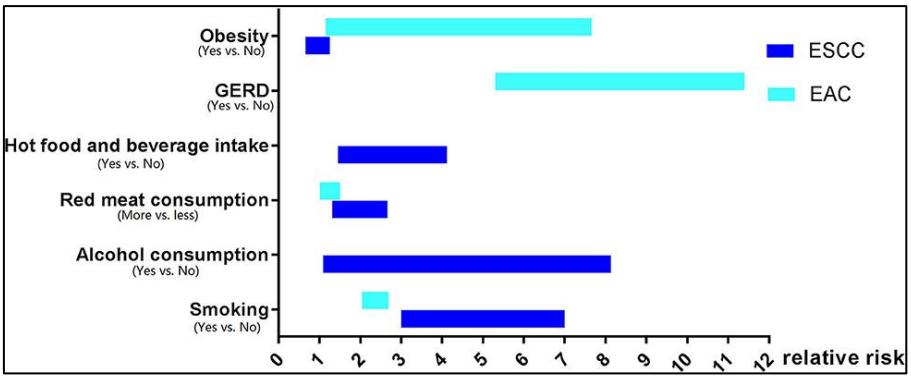


Figure 2.1: The Risk Factors for ESCC and EAC (Yang et al, 2020, p.1727).

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, analysing the differential gene expression able to obtain insight into the important underlying biological mechanisms and pathways of a particular disease or condition.

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

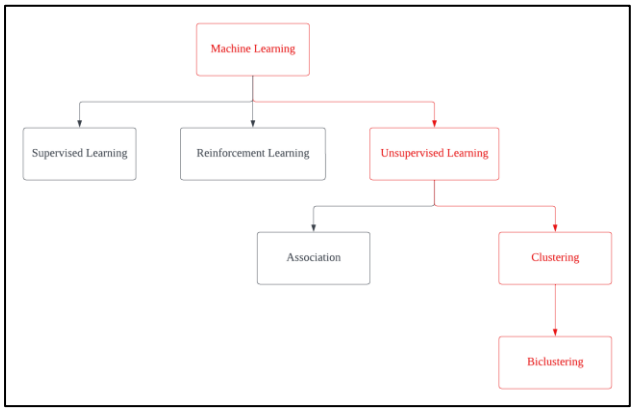


Figure 2.2: The Types of Machine Learning

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.

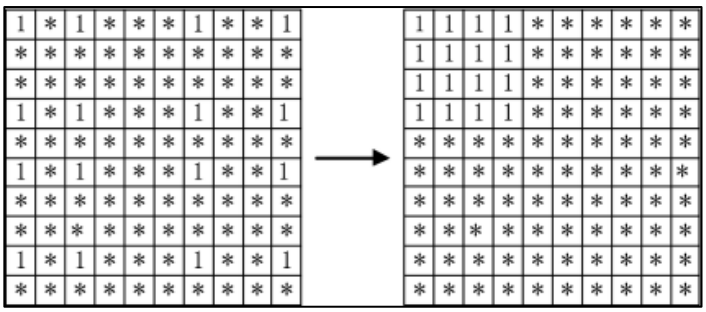


Figure 2.3: Biclusters of 1's in a Binary Matrix

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.

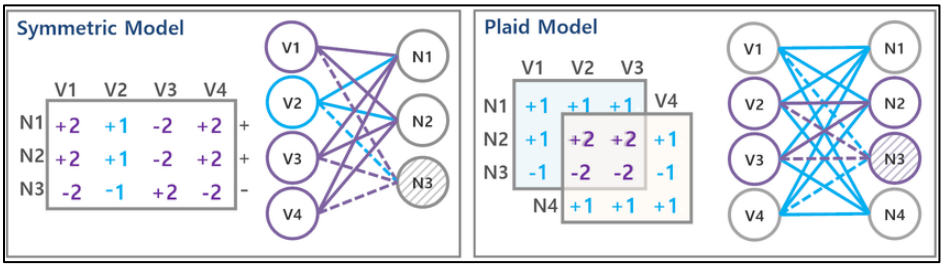


Figure 2.4: The Working Theory of Plaid Biclustering Model (Henriques and Madeira, 2015, pp 1-15)

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). The popular classification methods 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 when 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 computationally 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 had 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 methodologies were 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 considered 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 to accurately identify and gather possible biomarkers for esophageal cancer.

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

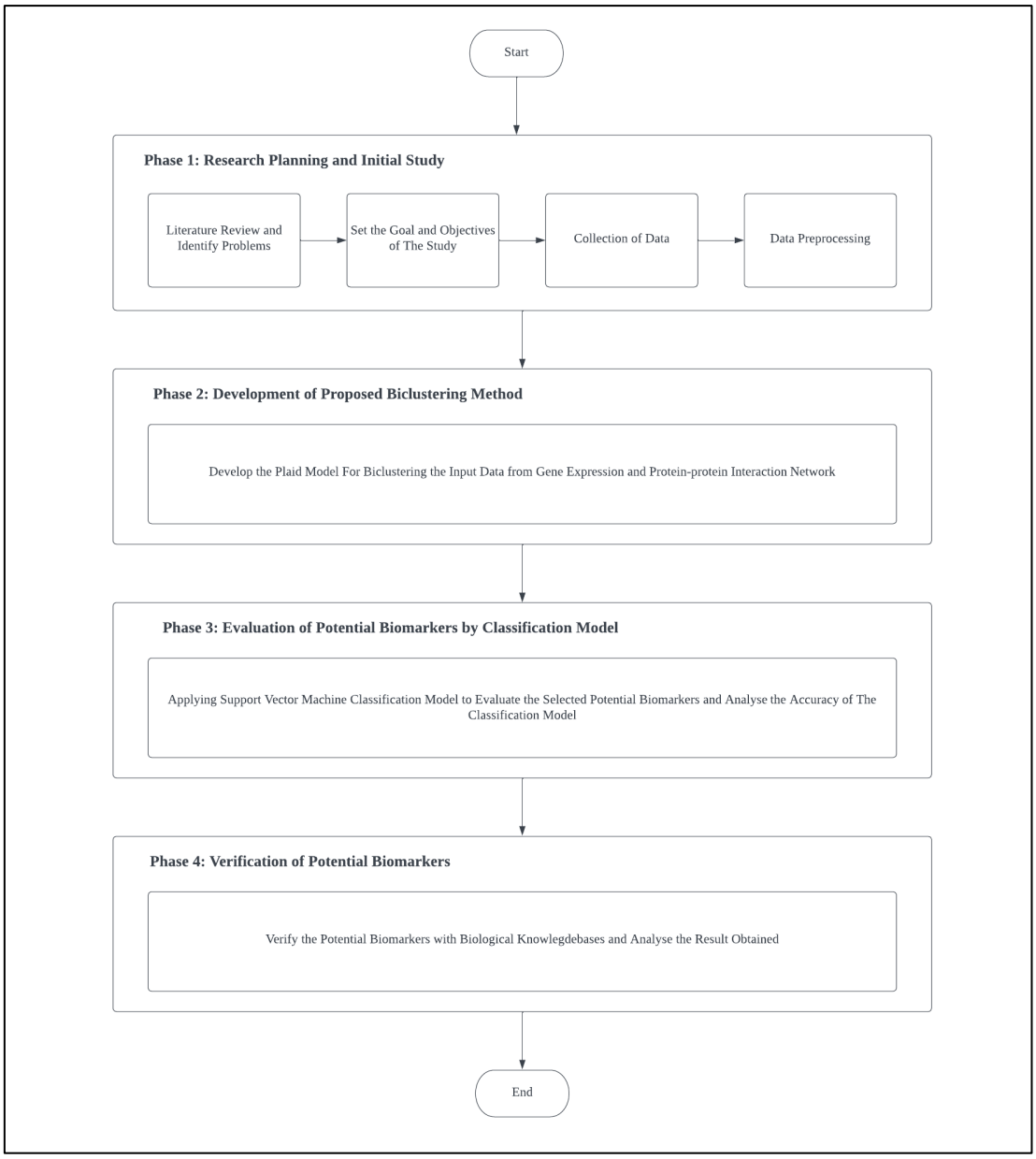


Figure 3.1: Research Framework

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 showed 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 was 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 aided in preparing for and carrying out of the author's own research and helped authors prevent duplicating previous studies. The author had a complete view of problem areas by analyzing previous issues-related work done by other researchers. This allowed 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. For futher clarification, plaid model allows data points to be the members of numerous biclusters with various intensities compared to other biclustering algorithm. 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. Support vector machine can be said as a modern machine learning method for efficiently classifying high-dimensional into a smaller group of datasets (Ozer et al, 2020). This process is an effective tool for classification tasks since it rapidly divides subgroups (Ozer et al, 2020).

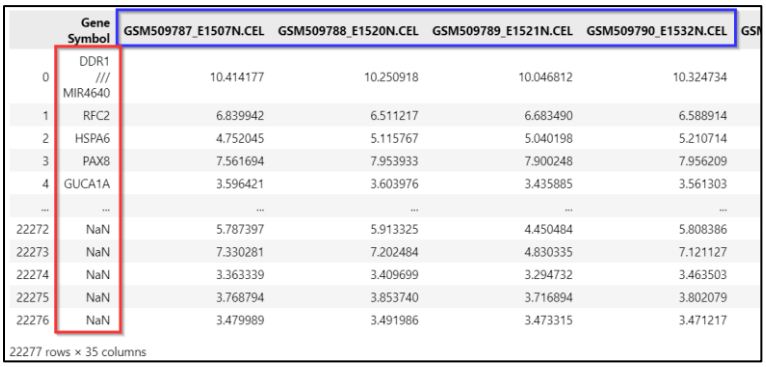
SVM performs well, particularly when dealing with situations where there are two distinct groups (Keerthana et al, 2023), like tumour samples and normal samples. The goal is to find the best boundaries to separate different data classes, increasing the distance between them while lowering classification mistakes (Keerthana et al, 2023). By using this procedure, support vector machines can efficiently categorise unobserved data points according to their underlying features (Keerthana et al, 2023).

3.2.4 Phase 4: Verification of Potential Biomarkers

During the classification process, the chosen biomarkers underwent training and testing. The biological knowledge base was subsequently 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 which can be found in the NCBI and UniProt. Researchers able to validate the biomarkers for a certain disease by searching the information of genes and related experiments that had been done previously. 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	GSM509790_E1532N.CEL	GSM
0	DDR1 /// MIR4640	10.414177	10.250918	10.046812	10.324734	
1	RFC2	6.839942	6.511217	6.683490	6.588914	
2	HSPA6	4.752045	5.115767	5.040198	5.210714	
3	PAX8	7.561694	7.953933	7.900248	7.956209	
4	GUCA1A	3.596421	3.603976	3.435885	3.561303	
...	...	...	...	...	...	
22272	NaN	5.787397	5.913325	4.450484	5.808386	
22273	NaN	7.330281	7.202484	4.830335	7.121127	
22274	NaN	3.363339	3.409699	3.294732	3.463503	
22275	NaN	3.768794	3.853740	3.716894	3.802079	
22276	NaN	3.479989	3.491986	3.473315	3.471217	

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

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.

Both datasets can be retrieved from below link:

- 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

3.4 The General Flow of Plaid Model

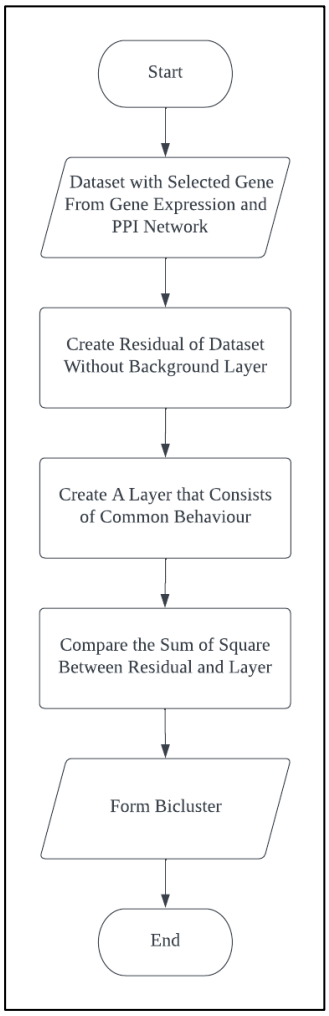


Figure 3.4: General Flow of Plaid Model

The basic concept of the plaid model is it formed a residual dataset based on the input data (Dataset with Selected Gene from Gene Expression and PPI Network) that had done the data preparation step. This residual dataset will exclude the background layer that underlying in the input data. Then, the algorithm will retrieve the rows and columns randomly from the residual to create a new layer that consists of the common behaviour. The significance of the layer will be tested by calculating the sum of square of data points between residual and layer. The data point considered as significant was retrieved out to form a bicluster. When there are four groups of biclusters is formed, then the process will be terminated.

3.5 Performance Measurement

In this research, the performance of the plaid biclustering algorithm was evaluated by the ten-fold cross validation and confusion matrix through the SVM classification. Ten runs of the experiments were carried out to obtain a more reliable measurement of the performance for the datasets. The dataset was divided into 60 to 40 percent of training set and testing set respectively with the stratified split of target variable. 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 to evaluate the effectiveness of a classification algorithm (Luque et al, 2019). 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 showed the confusion matrix. True Positives are the number of correctly predicted positive instances (Vujović, 2021). 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 (Vujović, 2021). 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) (Vujović, 2021). 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) (Vujović, 2021). Recall is the ratio of accurate positive predictions to all positive cases and the formula is TP/(TP+FN) (Vujović, 2021). According to Scikit Learn (n.d), F1 score is a measurement for evaluating the overall performance by providing the balance between of precision and recall and the formula is (2*precision*recall)/ (precision + recall).

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 the findings by undertaking a thorough search. The biological context validation stage ensured 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.

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

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.

Higher SSE values indicated greater dispersion and variability within clusters, potentially indicating insufficient clustering. In contrast, a low SSE score suggested less dispersion and variability within clusters implied that the data points in the dataset are more closely clustered.

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 method.

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 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.

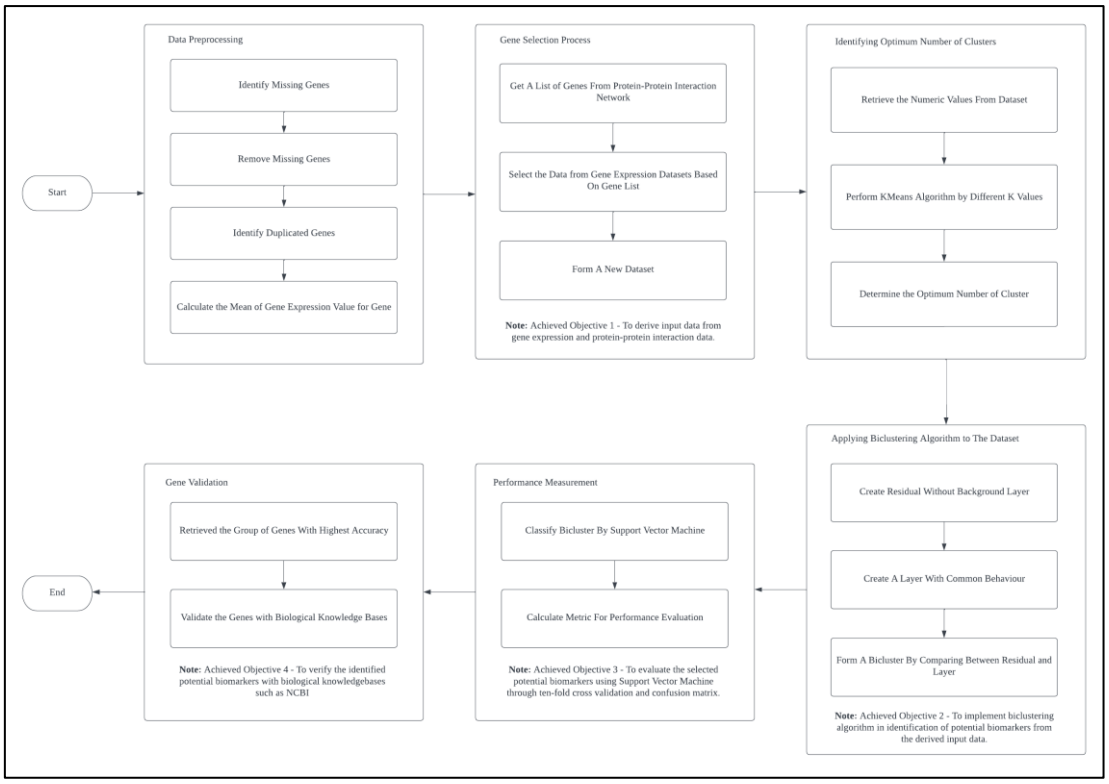


Figure 4.1: Development Process

4.2 Data Preparation

Data preparation is important to transform the dataset into an appropriate form for analysis and interpretation. There are two datasets used in this study. 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 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, human genes were 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. PPI data provided the full understanding of the connection between genes' activity. Filtering the gene expression dataset by the PPI data enabled to only focus on the important underlying patterns of gene expression dataset and indirectly enhanced the performance of the model to identify the possible biomarkers of EC.

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.

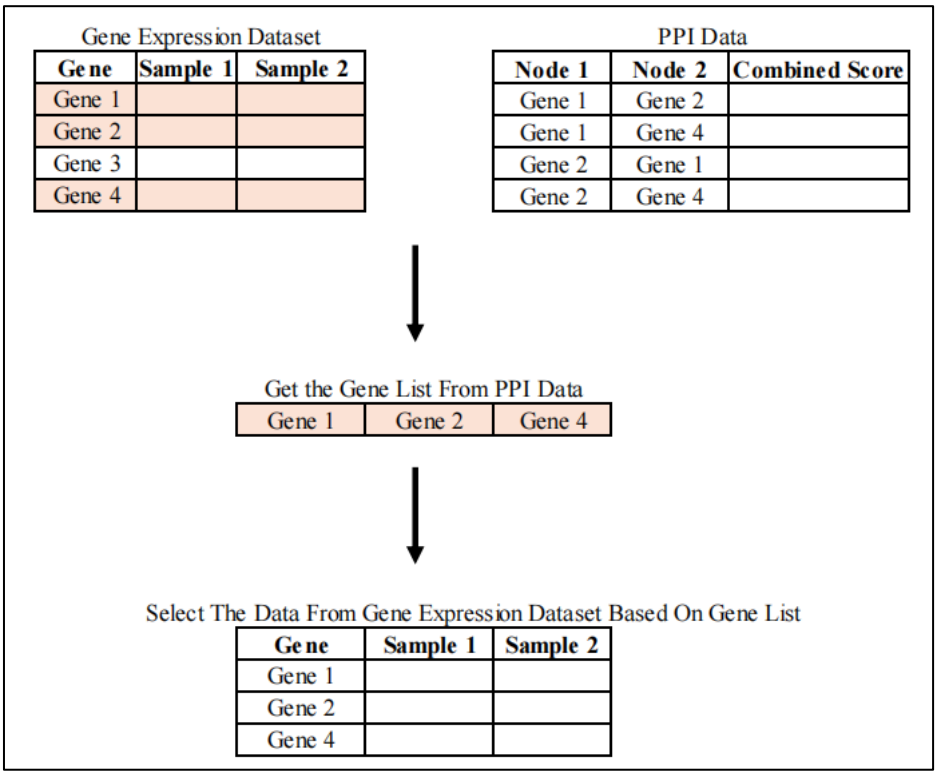


Figure 4.2: Gene Selection Working Process

	Gene Symbol	GSM509787_E1507N.CEL	GSM509788_E1520N.CEL	GSM509789_E1521N.CEL	GSM509790_E1532N.CEL
0	HSPA6	5.305487	5.608065	5.433042	5.556593
1	GUCA1A	4.165863	4.085270	3.955792	4.104240
2	CCL5	7.191147	7.422020	7.854530	6.542158
3	MMP14	6.839935	6.682461	6.629862	6.754795
4	TRADD	6.915792	6.779200	6.876954	7.094586

Figure 4.3: Example of 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. KMeans algorithm had been used in order to perform the elbow method. KMeans algorithm in elbow method is to group the data points according to the nearest distance. Then, the distances of a group of data were calculated for the sum of square error. Figure 4.4 demonstrated that the optimum number of clusters is four.

When a smaller number of clusters was used in the biclustering algorithm, the biclusters were losing the important features and patterns of the dataset. This is because there are less representations of the gene expression dataset to perform in the bicluster cause the interactions between the genes across the sample data to be ignored. However, the biclusters were capturing more noisier data when excessive number of clusters was used in the biclustering algorithm. The increasing of noisy data in the biclusters resulted in the difficulty to analyze and differentiate the important features from the irrelevant data. In conclusion, identify the optimum number of clusters is important to help the biclustering algorithm in capturing the gene expression patterns of the data.

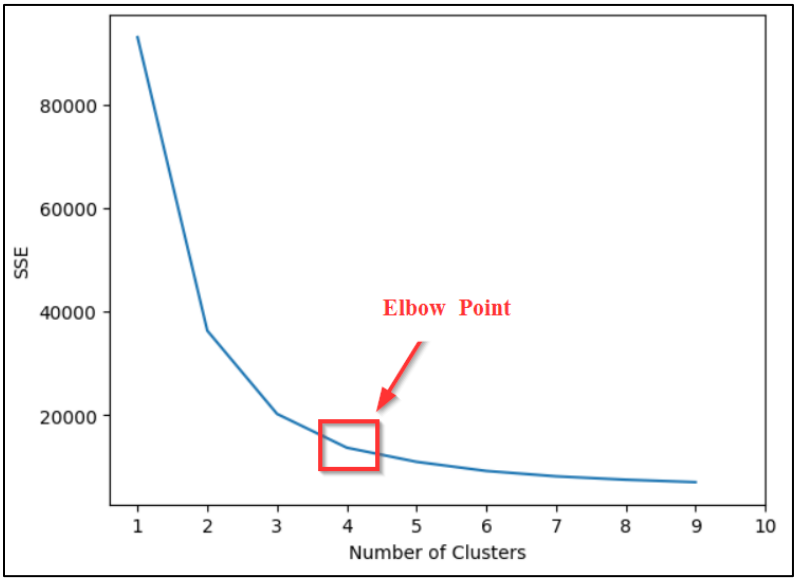


Figure 4.4: Optimum Number of Cluster by Using Elbow Method

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.

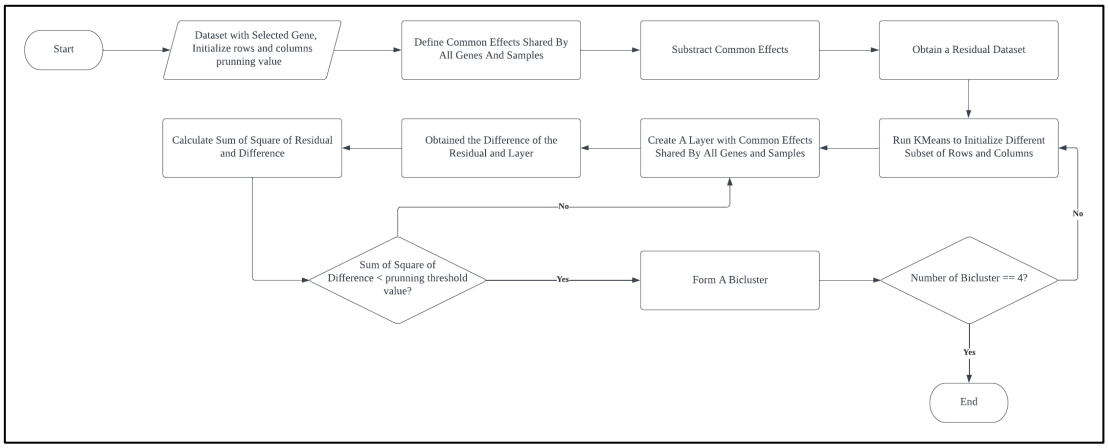


Figure 4.5: Basic Architecture of Plaid Biclustering

4.4.1 Create Background Layer from Dataset with Selected Gene for Pattern Capture

There is a background layer in the Plaid bicluster model. Background layers in Plaid models indicated common effects shared by all genes and samples. In this step, the mean, row effects and column effects of the dataset with selected gene will be calculated. This method captures both the overall average behavior of the row and column divergent by computing row and column effects. By forming the new layers, particular effects can be separated from background layers to show biclusters that are specific to a condition or treatment.

4.4.2 Substract Background Layer/Common Effects

By subtracting the background layer from the dataset with selected gene, the algorithm effectively removed the common impact represented by the background

layer from the dataset of selected gene. This procedure updates the dataset of selected genes to concentrate on any remaining precise changes of the genes' activity across sample data.

4.4.3 Formed A Collection of Bicluster

The process required in finding coherent groups of genes and samples, capturing their shared behavior in a common layer, verifying that this behavior is meaningful, and finally classifying these groups as biclusters if specific criteria are met. Using this method, important features were obtained.

4.4.3.1 Run K-Means to Initialize Rows and Columns

K-Means algorithm is a popular partition method to define the dissimilarity between the points (Sinaga and Yang, 2020). As a result, expression patterns in genes and samples by using K-Means algorithm were comparable. These expression patterns showed genes that are co-regulated in specific situations or samples that react to stimuli in a comparable way. In this study, the K-Means algorithm was used to divide rows and columns, effectively breaking up the dataset into smaller parts with similar features. At this stage, the rows and columns for the new layer can be initialized.

4.4.3.2 Create A Layer with Common Effects Shared by All Genes and Samples

After the rows and columns are initialized, a new layer was formed. This layer represented the common effects of all genes and samples in the dataset. It was constructed by averaging the residuals and combining row and column effects. To elaborate, averaging the residuals was to determine the mean value of the entire dataset. The row effect reflected the mean value of each row, whereas the column effect indicated the mean value of each column. Essentially, this layer provides the overall

behavior observed in the subset of data determined by the clusters created in the previous stage.

4.4.3.3 Sum of Square

In order to ensure that the common behaviors observe were meaningful rather than just random occurrences, the variants in the residual (after subtracting common effect layer by dataset with selected gene) to those captured by the common layer had been compared. The difference of residual and layer was calculated. Then, the differences and the residual were calculated to obtain their sum of square value. If the pruning threshold value is higher than the sum of square of the differences, it means that the layer does capture enough diversity in the data and is thus significant.

4.4.3.4 Form Biclusters

After the significant data points been selected, a bicluster layer is formed. A collection of genes and samples that showed similar behavior or patterns in the dataset were represented by this layer. The bicluster layer was then added to the layer list, and the procedure continue to search for further bicluster. Figures below show the example of the bicluster that had been formed. A total of four biclusters were formed in this study.

Gene Symbol	TMEM39A	NUP107	TMEM38B	PBK	ASPM	NELFCD	DNMT3B	TBL1XR1	GINS2	COL5A2
GSM509804_E1507T.CEL	6.611383	9.726017	5.445221	7.313210	7.352215	8.965812	6.870785	6.893274	7.324547	9.898811
GSM509809_E1542T.CEL	6.389347	8.623677	4.765128	7.265171	8.088150	8.763703	5.790180	5.530707	7.380882	8.929217
GSM509810_E1546T.CEL	6.724950	8.185211	5.477874	8.210063	7.910408	9.619201	6.787680	7.152694	7.778961	10.010738
GSM509812_E1584T.CEL	6.587234	8.930046	4.597372	8.368051	8.356823	8.400663	6.114453	6.503139	8.567287	8.953696
GSM509813_E1589T.CEL	6.583535	9.160049	6.840389	8.215945	8.576583	9.793823	8.950951	7.235061	9.331082	9.496884

Figure 4.6: Bicluster 1

Gene Symbol	DENND1B	RNF39	SLC27A6	GIPC2	EPB41L4A	ADAMTSL4	RIPK4	UBAP1	VPS37B	CSNK1E
GSM509803_E2644N.CEL	9.025555	8.580828	4.229821	5.761416	5.770840	5.621589	10.616623	8.803706	9.253934	7.787345
GSM509804_E1507T.CEL	8.202442	6.716919	4.103199	3.446680	4.413280	5.258764	9.521838	7.682166	6.960880	7.428256
GSM509805_E1520T.CEL	7.325726	5.681840	3.881914	3.485505	4.465795	5.194590	7.898191	7.109491	7.278069	7.222471
GSM509806_E1521T.CEL	8.159822	5.902807	3.941313	3.378748	4.152978	5.104112	8.974263	7.553683	8.003548	6.968401
GSM509809_E1542T.CEL	8.361134	6.493334	4.079763	3.801562	4.456992	5.741768	8.988443	7.181832	8.338481	7.648359

Figure 4.7: Bicluster 2

Gene Symbol	DVL3	EPHB4	APOC1	LAMB3	HOXD11	ANO1
GSM509819_E1796T.CEL	7.691792	7.516879	8.240716	10.040889	4.567577	4.450293

Figure 4.8: Bicluster 3

Gene Symbol	NREP	HEXB	EPHB4	KIF3B	BRCA1	SAP30	PTK2	LAMB3	MBD4	CHN1	ALMS1	HOXD11	BANP
GSM509816_E1614T.CEL	7.82422	8.918272	7.471044	6.550222	6.12884	5.130646	9.079471	9.57014	8.38678	6.323455	6.883904	5.71802	7.740366

Figure 4.9: Bicluster 4

4.5 Performance Measurements

A SVM classifier was used to discover potential EC cancer biomarkers from the collection of biclusters. However, the biclustering result indicated that the retrieved sample is cancerous. The gene expression dataset used for biclustering algorithm after the data preprocessing and the gene selection process consisted with 17 normal samples and 17 cancerous samples. This balanced dataset suggested that the bias of the biclustering algorithm toward cancer cases can be denied. The plaid model was to discover the key features of the patterns of genes. Hence, the resulting biclusters that only consisted of the cancerous samples data is due to the gene expression values that presented in the data were showing stronger expression patterns than normal samples. The figures below showed the bicluster data with the 'Target' class.

Gene Symbol	NUP107	TMEM38B	PBK	ASPM	NELFCD	DNMT3B	TBL1XR1	GINS2	COL5A2	Target
GSM509804_E1507T.CEL	9.726017	5.445221	7.313210	7.352215	8.965812	6.870785	6.893274	7.324547	9.898811	1
GSM509809_E1542T.CEL	8.623677	4.765128	7.265171	8.088150	8.763703	5.790180	5.530707	7.380882	8.929217	1
GSM509810_E1546T.CEL	8.185211	5.477874	8.210063	7.910408	9.619201	6.787680	7.152694	7.778961	10.010738	1
GSM509812_E1584T.CEL	8.930046	4.597372	8.368051	8.356823	8.400663	6.114453	6.503139	8.567287	8.953696	1
GSM509813_E1589T.CEL	9.160049	6.840389	8.215945	8.576583	9.793823	8.950951	7.235061	9.331082	9.496884	1
GSM509815_E1610T.CEL	10.200163	6.290956	7.325244	9.338510	9.725768	6.838957	8.715199	9.070411	8.867885	1
GSM509816_E1614T.CEL	9.186521	5.895004	8.165070	9.464792	8.760688	6.958920	7.352432	8.375379	10.165669	1
GSM509817_E1635T.CEL	9.105222	5.922400	8.535196	8.996614	9.897316	6.844505	7.660548	8.854069	11.291970	1
GSM509818_E1709T.CEL	9.104914	5.666086	7.930660	8.449101	9.843952	7.864205	6.703331	8.901941	9.793506	1
GSM509819_E1796T.CEL	9.671777	6.246486	7.543486	9.275292	9.300892	7.719712	6.968967	8.497313	9.826365	1

Figure 4.10: Example of Bicluster 1 with Target Class

Gene Symbol	RNF39	SLC27A6	GIPC2	EPB41L4A	ADAMTSL4	RIPK4	UBAP1	VPS37B	CSNK1E	Target
GSM509803_E2644N.CEL	8.580828	4.229821	5.761416	5.770840	5.621589	10.616623	8.803706	9.253934	7.787345	0
GSM509804_E1507T.CEL	6.716919	4.103199	3.446680	4.413280	5.258764	9.521838	7.682166	6.960880	7.428256	1
GSM509805_E1520T.CEL	5.681840	3.881914	3.485505	4.465795	5.194590	7.898191	7.109491	7.278069	7.222471	1
GSM509806_E1521T.CEL	5.902807	3.941313	3.378748	4.152978	5.104112	8.974263	7.553683	8.003548	6.968401	1
GSM509809_E1542T.CEL	6.493334	4.079763	3.801562	4.456992	5.741768	8.988443	7.181832	8.338481	7.648359	1
GSM509810_E1546T.CEL	6.252399	3.730487	3.906003	4.328748	5.218253	9.324331	7.611860	6.452536	7.171389	1
GSM509811_E1566T.CEL	5.428539	3.798422	3.706377	4.186752	4.925162	8.660770	7.062897	6.387506	7.272552	1
GSM509812_E1584T.CEL	5.950275	3.698107	3.992634	4.286096	5.046298	9.978480	7.064077	7.659996	7.070365	1
GSM509814_E1603T.CEL	5.662010	3.787020	4.987507	4.451226	4.990375	10.032413	8.168227	8.026121	7.377919	1
GSM509815_E1610T.CEL	5.830913	4.241176	3.991872	3.721478	4.925162	8.234292	6.871717	6.611458	6.552709	1
GSM509816_E1614T.CEL	6.049930	3.574875	3.648740	4.196909	5.024340	8.930225	7.779759	7.651306	7.266086	1
GSM509817_E1635T.CEL	6.446707	3.693587	3.531805	4.062639	5.198883	8.526820	8.272057	7.553935	6.904186	1
GSM509818_E1709T.CEL	6.393301	3.741121	4.004577	4.267456	4.916629	8.669479	7.874860	7.918919	6.861825	1
GSM509819_E1796T.CEL	5.434447	3.656964	5.455840	4.181800	5.377791	7.570866	7.222414	7.007866	6.941757	1

Figure 4.11: Example of Bicluster 2 with Target Class

Gene Symbol	DVL3	EPHB4	APOC1	LAMB3	HOXD11	ANO1	Target
GSM509819_E1796T.CEL	7.691792	7.516879	8.240716	10.040889	4.567577	4.450293	1

Figure 4.12: Bicluster 3 with Target Class

Gene Symbol	NREP	HEXB	EPHB4	KIF3B	BRCA1	SAP30	PTK2	LAMB3	MBD4	CHN1	ALMS1	HOXD11	BANP	Target
GSM509816_E1614T.CEL	7.82422	8.918272	7.471044	6.550222	6.12884	5.130646	9.079471	9.57014	8.38678	6.323455	6.883904	5.71802	7.740366	1

Figure 4.13: Bicluster 4 with Target Class

4.5.1 Prepare the Gene Expression Dataset for Classification

Since the biclusters only consisted of cancerous samples, hence the data cannot undergo the classification directly. This is due to the classification required to leam the data between different targets. To enable the biclusters for the classification purpose, a few of steps had been carried out for the determination of model's

performance. Figure 4.14 showed the flow on classifying the genes that had been extracted from bicluster.

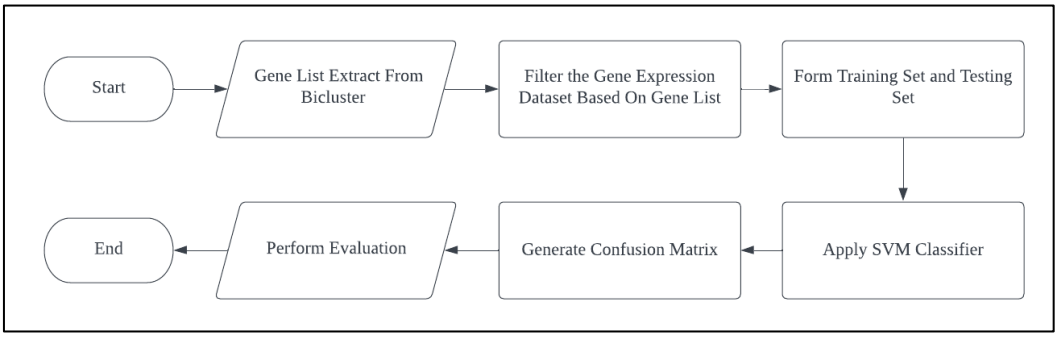


Figure 4.14: The Flow of Classification

The goal of biclustering algorithms was to discover key features by showing patterns in gene and sample data. Thus, the genes found inside the bicluster were important indicators that predictive of EC cancer. It indicated that these genes exhibit patterns that implied their involvement in disease. Hence, the genes that found inside the bicluster were used to filter the Gene Expression Dataset. Furthermore, some genes appearred in multiple bicluster which means these genes were playing an important role in the development of EC. To improve classification results, these genes that occurred in more than one bicluster were used to filter the Gene Expression Dataset for another classification process.

In summary, three Gene Expression Datasets were used to develop SVM classifiers. The datasets retrieved for classification were illustrated in the images below. Figure 4.16 displayed the Gene Expression Dataset which only included genes from biclusters and consisted of 34 samples and 285 genes. Figure 4.17 represented the Gene Expression Dataset which contained genes that occurred in more than one bicluster and consisted of 34 samples and 3 genes. Meanwhile, the Original Gene Expression Dataset which included 2735 genes, and 34 samples was displayed in Figure 4.18.

Gene Symbol	NELFCD	DNMT3B	RIPK4	TBL1XR1	UBAP1	GINS2	VPS37B	COL5A2	CSNK1E	Target
GSM509787_E1507N.CEL	8.847780	5.429582	10.528422	6.103538	8.576635	6.942437	9.204626	5.803623	7.330217	0
GSM509788_E1520N.CEL	8.486673	5.233733	10.102096	5.805960	8.265671	6.760712	7.639002	5.855633	7.400418	0
GSM509789_E1521N.CEL	8.597308	5.480147	10.313087	5.571494	8.563540	6.950697	8.970822	6.742893	7.682231	0
GSM509790_E1532N.CEL	8.602311	5.228284	10.123108	4.935344	8.577154	6.846777	9.346425	5.990346	7.277551	0
GSM509791_E1535N.CEL	8.247174	5.252625	10.470718	4.991667	8.115302	6.561701	8.833978	5.726995	7.696482	0

Figure 4.15: Example of Gene Expression Dataset that Involved Genes Extracted from Biclusters

Gene Symbol	EPHB4	LAMB3	HOXD11	Target
GSM509787_E1507N.CEL	6.892630	8.098309	4.216160	0
GSM509788_E1520N.CEL	6.778301	6.880918	4.863400	0
GSM509789_E1521N.CEL	7.016927	8.011573	4.607442	0
GSM509790_E1532N.CEL	6.984915	7.639520	4.455296	0
GSM509791_E1535N.CEL	7.114578	7.014760	4.367265	0

Figure 4.16: Example of Gene Expression Dataset that the Genes that Occurred in Multiple Biclusters

Gene Symbol	TBC1D2	B4GALT7	CASP8AP2	NACA2	TRDN	SCAF4	LAMA1	FBXO31	SLC44A1	Target
GSM509787_E1507N.CEL	7.765013	5.949763	7.054392	3.451194	3.389328	6.060083	4.591470	6.167321	6.157059	0
GSM509788_E1520N.CEL	7.297379	5.828337	6.838957	3.677271	3.281392	5.524395	4.637121	6.635602	6.376391	0
GSM509789_E1521N.CEL	7.498661	5.811177	7.229369	3.575415	3.498269	6.353812	4.658779	6.359670	5.889888	0
GSM509790_E1532N.CEL	7.672872	5.988737	7.151789	3.647739	3.266002	6.034053	4.482420	6.408367	6.167377	0
GSM509791_E1535N.CEL	7.479567	5.652972	6.965022	3.517448	3.465066	6.884892	4.618102	6.108703	6.840216	0

Figure 4.17: Example of Original Gene Expression Dataset

4.5.2 Apply SVM Classifier to the Gene Expression Dataset

The dataset was split into features and the target variable in order to apply the SVM classifier. The final column named “Target” identified as the target variable while other columns were chosen as features. Subsequently, the dataset was divided into training and test sets, with a 60 percent training to 40 percent test ratio. The performance of a linear SVM classifier was evaluated using a ten-fold cross validation technique. Cross-validation scores were computed, and predicted labels were obtained for each fold. A confusion matrix was also created to evaluate the classifier's performance. Furthermore, different performance indicators, such as accuracy, precision, recall, specificity, and F1 score, were used to evaluate the classifier's effectiveness. Tables below show the performance of the SVM classifier for each Gene Expression Dataset based on ten-fold cross validation and confusion matrix.

Table 4.1: Performance Evaluation of Gene Expression Dataset Based on 10-Fold Cross Validation

	Gene Expression Dataset That Involved Genes		
	In All Bicluster	Occur In More Than One Bicluster	Original Dataset
10-Fold Cross Validation			
Fold 1	1	1	1
Fold 2	1	1	1
Fold 3	1	0.75	1
Fold 4	1	1	1
Fold 5	1	1	1
Fold 6	0.6667	1	0.6667
Fold 7	1	1	1
Fold 8	1	1	1
Fold 9	1	1	1
Fold 10	1	1	1
Average	0.9667	0.975	0.9667

Table 4.2: Performance Measurement of Gene Expression Dataset Based on Confusion Matrix

Metrics	Gene Expression Dataset That Involved Genes		
	In All Bicluster	Occur In More Than One Bicluster	Original Dataset
Accuracy	0.9706	0.9706	0.9706
Precision	1	0.9444	1
Recall	0.9412	1	0.9412
Specificity	1	0.9412	1
F1 Score	0.9697	0.9714	0.9697

4.5.3 Verify the Selected Potential Biomarkers

As the classification accuracy for three datasets was almost the same, a t test with significance level of 0.05 was conducted on the accuracy values obtained from multiple runs to analyze the relationship of three datasets. The results of the t-test were shown below.

```
T-test between Genes In All Biclusters and Genes Occur in Multiple Biclusters:
T-statistic: 0.8847, P-value: 0.3880


T-test between Genes In All Biclusters and Original Dataset:
T-statistic: 0.0000, P-value: 1.0000


T-test between Genes Occur in Multiple Biclusters and Original Dataset:
T-statistic: -0.8847, P-value: 0.3880
```

Figure 4.18: T-Test Result

The result of SVM classifier and t-test indicated that the gene expression dataset that involved genes that occurred in more than one biclusters achieved the better result. Hence, the genes in this dataset were further validated 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 better classification accuracy when SVM classification model is applied to different Gene Expression Dataset. 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

RESULT DISCUSSION

5.1 Input Data from Gene Expression Dataset and PPI Network

The PPI network provided an insight of the interaction between genes which improved the understanding of the underlying biological process within gene expression dataset. Hence, by focusing on the genes that found in the PPI network enabled to focus on the important features or meaningful patterns of genes in gene expression dataset. Therefore, the genes in the PPI network were extracted to filter out the noisy and irrelevant data such as the genes lacking biological interaction within the gene expression dataset. The filtered dataset formed a new input data which was represented in the form of Gene Expression. The gene selection process ensured that the biclustering algorithm and the analysis process highlighted the biological relevant genes which indirectly enhancing the accuracy of the model's performance. Figure 5.1 illustrated the input data that was used in the biclustering algorithm.

	Gene Symbol	GSM509787_E1507N.CEL	GSM509788_E1520N.CEL	GSM509789_E1521N.CEL	GSM509790_E1532N.CEL
0	HSPA6	5.305487	5.608065	5.433042	5.556593
1	GUCA1A	4.165863	4.085270	3.955792	4.104240
2	CCL5	7.191147	7.422020	7.854530	6.542158
3	MMP14	6.839935	6.682461	6.629862	6.754795
4	TRADD	6.915792	6.779200	6.876954	7.094586

Figure 5.1: Gene Expression Dataset after Filtering with the PPI Network

The dimension of the dataset was 2735 genes across 34 samples. Each row represented a gene symbol (red box) while each column represented a sample (blue box). The value within the Gene Expression Dataset indicates the gene expression value of the respective genes across the sample (green box). In a nutshell, the dataset provided a comprehensive insight of the patterns of genes across different samples.

5.2 The Involvement of PPI data in Gene Expression Dataset

Table 5.1: The Accuracy of Gene Expression Dataset with and without PPI data

Test Size	Accuracy	
	Gene Expression Dataset	Gene Expression Dataset with PPI Data
0.1	1	1
0.2	0.8571	1
0.3	1	0.9091
0.4	0.9286	0.9286
Average	0.9464	0.9594

Table 5.1 compared the accuracy of SVM classifier using two datasets of varying test sizes to evaluate the effectiveness of the involvement of PPI data with gene expression dataset. According to the average value of performance, Gene Expression Dataset with PPI achieved 95.94 percent accuracy while Gene Expression Dataset indicated 94.64 percent accuracy. The results suggested that using PPI data enhanced the classification performance of the SVM model, demonstrating its potential value in detecting EC cancer biomarkers. Besides that, the higher accuracy achieved by the involvement of PPI data suggested that the importance of combining multiple biological relevance data was allowing to obtain a more comprehensive biological insights that underlying the data and indirectly leading to more accurate biomarker identification.

5.3 Plaid Biclustering Algorithm in Identifying the Biclusters

The plaid biclustering algorithm was applied to the newly generated input data. The biclustering algorithm's capability to recognize similar expression patterns leads the grouping of genes and samples into biclusters. The plaid model identified subset of genes with comparable expression across samples. In detail, the plaid model clustered genes that showed similar patterns of expression under certain conditions. Each bicluster represented a distinct set of genes and samples that shared common

function under those conditions. Besides that, the groups of genes also highlighted the behavior or responses across the sample to a related environment.

A total of four biclusters were. In the dataset produced by the plaid biclustering model, rows represented samples while columns represented genes. The dimension of the Bicluster 1 is 124 genes across 10 samples. The dimension of the Bicluster 2 is 145 genes across 14 samples. The dimension of the Bicluster 3 is 6 genes across 1 sample while the dimension of the Bicluster 4 is 13 genes across 1 sample.

Gene Symbol	TMEM39A	NUP107	TMEM38B	PBK	ASPM	NELFCD	DNMT3B	TBL1XR1	GINS2	COL5A2
GSM509804_E1507T.CEL	6.611383	9.726017	5.445221	7.313210	7.352215	8.965812	6.870785	6.893274	7.324547	9.898811
GSM509809_E1542T.CEL	6.389347	8.623677	4.765128	7.265171	8.088150	8.763703	5.790180	5.530707	7.380882	8.929217
GSM509810_E1546T.CEL	6.724950	8.185211	5.477874	8.210063	7.910408	9.619201	6.787680	7.152694	7.778961	10.010738
GSM509812_E1584T.CEL	6.587234	8.930046	4.597372	8.368051	8.356823	8.400663	6.114453	6.503139	8.567287	8.953696
GSM509813_E1589T.CEL	6.583535	9.160049	6.840389	8.215945	8.576583	9.793823	8.950951	7.235061	9.331082	9.496884

Figure 5.2: Example of Bicluster 1 after Implement Plaid Biclustering Model

Gene Symbol	DENND1B	RNF39	SLC27A6	GIPC2	EPB41L4A	ADAMTSL4	RIPK4	UBAP1	VPS37B	CSNK1E
GSM509803_E2644N.CEL	9.025555	8.580828	4.229821	5.761416	5.770840	5.621589	10.616623	8.803706	9.253934	7.787345
GSM509804_E1507T.CEL	8.202442	6.716919	4.103199	3.446680	4.413280	5.258764	9.521838	7.682166	6.960880	7.428256
GSM509805_E1520T.CEL	7.325726	5.681840	3.881914	3.485505	4.465795	5.194590	7.898191	7.109491	7.278069	7.222471
GSM509806_E1521T.CEL	8.159822	5.902807	3.941313	3.378748	4.152978	5.104112	8.974263	7.553683	8.003548	6.968401
GSM509809_E1542T.CEL	8.361134	6.493334	4.079763	3.801562	4.456992	5.741768	8.988443	7.181832	8.338481	7.648359

Figure 5.3: Example of Bicluster 2 after Implement Plaid Biclustering Model

Gene Symbol	DVL3	EPHB4	APOC1	LAMB3	HOXD11	ANO1
GSM509819_E1796T.CEL	7.691792	7.516879	8.240716	10.040889	4.567577	4.450293

Figure 5.4: Bicluster 3 after Implement Plaid Biclustering Model

Gene Symbol	NREP	HEXB	EPHB4	KIF3B	BRCA1	SAP30	PTK2	LAMB3	MBD4	CHN1	ALMS1	HOXD11	BANP
GSM509816_E1614T.CEL	7.82422	8.918272	7.471044	6.550222	6.12884	5.130646	9.079471	9.57014	8.38678	6.323455	6.883904	5.71802	7.740366

Figure 5.5: Bicluster 4 after Implement Plaid Biclustering Model

5.4 Applying SVM Classifier for the Performance Evaluation

The samples retrieved by the plaid biclustering model showed that all the samples retrieved were identified as tumor. Since there is only one class in the target column, it indicated that the biclustering technique successfully identified the groups of genes that exhibit similar expression features to the EC. As mentioned, biclustering

algorithm grouped genes based on their expression patterns. Thus, the genes within the resulted biclusters can be considered as the possible indicators in the development of EC due to the identified samples are tumors. As a result, the genes identified within the biclusters were used to filter the data from the original gene expression dataset. This gene selection process reduced the dataset by focusing on the genes that were relevant to EC. For further classification purpose, three different datasets were compared which are gene expression dataset that involved genes in all bicluster, gene expression dataset that involved genes occurred in more than one bicluster and original gene expression dataset respectively. Figure below illustrated the gene selection process on the biclusters to filter the data from gene expression dataset.

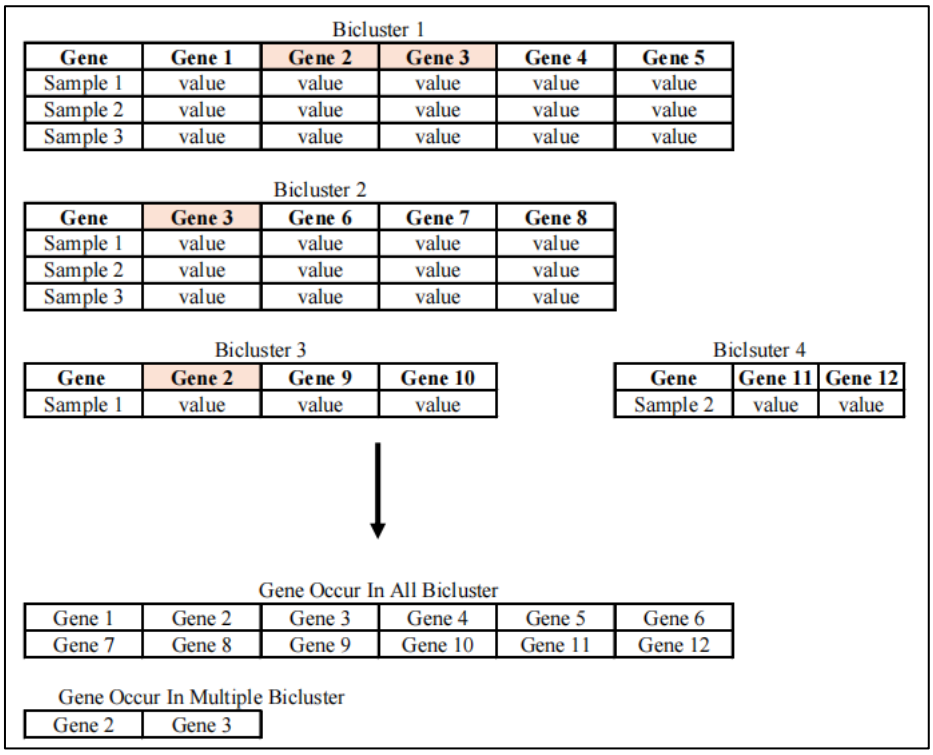


Figure 5.6: The Gene Selection Process Done on Biclusters

5.4.1 Determining the Optimum Train Test Split Ratio

Before evaluating the performance of the SVM on three datasets, the optimum train test split ratio was determined by using the original gene expression dataset. In this step, the accuracy value of each test size was evaluated to find the most effective ratio to split the dataset into training and testing set. This step is to ensure that the SVM

able to capture enough figure of the data and generalize effectively to the unseen data. Table 5.2 demonstrated the accuracy of the Gene Expression Dataset after the gene selection process across multiple runs with varying test sizes and different random states starting from 5 to 50.

Table 5.2: Accuracy of Gene Expression Dataset with Random State and Test Size

Run	Accuracy with Different Test Size			
	0.1	0.2	0.3	0.4
1	1	1	1	1
2	0.75	0.8571	0.9091	0.9286
3	0.75	0.8571	0.9091	0.9286
4	1	1	1	1
5	0.75	1	1	1
6	1	0.9571	0.9091	0.9286
7	1	1	1	1
8	1	0.8571	0.9091	0.9286
9	1	1	1	1
10	0.75	1	1	0.9286
Average	0.9	0.9428	0.9636	0.9643

Among the various test sizes, the 0.4 test size achieved better results than other test sizes. The 0.1 test size, 0.2 test size and 0.4 test size achieved 90 percent accuracy, 94.28 percent accuracy and 96.36 percent accuracy respectively. The 0.4 test size was chosen since it performed better than other test sizes. By assigning 40 percent of the data for testing, the model able to analyse more thoroughly across a larger percentage of the gene expression dataset and producing a more reliable and robust results. This test size also provided a balance by providing sufficient data for training the model while ensuring sufficient testing data for performance evaluation and validation. In conclusion, using 60 percent of training set and 40 percent of testing set enabled the model to capture enough pattern of the data and generalize the unseen data accurately.

5.4.2 Performance Evaluation of Gene Expression Dataset

Ten-fold cross validation was implemented to measure the performance of three distinct gene expression datasets. This validation provided the score value which indicated the overall performance of the model to each dataset. Ten-fold cross validation divided the dataset into ten equal sized folds. Then, the model was trained on the nine folds and tested on the remaining fold. The procedure resulted in a more accurate and reliable measurement as the bias of the performance had been avoided.

Table 5.3 indicated all three datasets achieved an accuracy of 100 percent in nine out of ten-fold. For the gene expression dataset that involved genes in all bicluster and original gene expression dataset achieved approximately 67 percent in fold six while the gene expression dataset that involved genes that occurred in more than one bicluster achieved 75 percent in fold three.

Table 5.3: Performance Evaluation Based on 10-Fold Cross Validation

	Gene Expression Dataset that Involved Genes		
	In All Bicluster	Occur In More Than One Bicluster	Original Dataset
Fold 1	1	1	1
Fold 2	1	1	1
Fold 3	1	0.75	1
Fold 4	1	1	1
Fold 5	1	1	1
Fold 6	0.6667	1	0.6667
Fold 7	1	1	1
Fold 8	1	1	1
Fold 9	1	1	1
Fold 10	1	1	1
Average	0.9667	0.975	0.9667

Table 5.4: Confusion Matrix Result based on Ten-Fold Cross Validation

Metrics	Gene Expression Dataset That Involved Genes		
	In All Bicluster	Occur In More Than One Bicluster	Original Dataset
No. of Features	283 genes	3 genes	2735 genes
Accuracy	0.9706	0.9705	0.9706
Precision	1	0.9444	1
Recall	0.9412	1	0.9412
Specificity	1	0.9412	1
F1 Score	0.9697	0.9714	0.9697

According to Cherradi et al (2021), the 100 percent accuracy in k-fold cross validation is due to the effectiveness of machine learning model to learn the patterns and relationships in the data. Besides that, feature selection also improved the ability of model to generalize the data (Cheraddi et al, 2021). Furthermore, the efficient feature extraction allowed the 100 percent accuracy performance to classify the fundus image into different glaucoma conditions through the 5-fold cross validation (Fuadah et al, 2022). Hence, the effectiveness of the model to learn the underlying patterns of data had the potential to increase the performance as 100 percent. In conclusion, the biclustering algorithm improved the performance of the model by comprehending the data completely.

Based on the results of the ten-fold cross validation, the accuracy varied across different folds. As there were 34 samples presented in the dataset, the training set for each fold was 30 or 31 samples while the testing set for each fold was 4 or 3 samples respectively. Hence, the variation in the ten-fold cross validation implied that the model's performance was influenced by the specific random splits of data used in each fold. This is because different subsets of the data were used for training and testing which indicated that the model was trained with diverse samples. Additionally, the gene expression dataset consists of genes from different biclusters which raised the issues regarding the strength and relevance of each gene to EC. Hence, the sudden

drop of accuracy at fold 6 and fold 3 was due to the exhibit higher variability of the expression patterns making the model unable to generalize well the data.

To overcome the issue of variation in accuracy value, the SVM classifier was ran ten times with different random state values to ensure that the model's performance is not excessively dependent on a certain random split of the data.

Table 5.5: Performance Evaluation Based on Multiple Run

Run	Gene Expression Dataset that Involved Genes		
	In All Bicluster	Occur In More Than One Bicluster	Original Dataset
1	1	1	1
2	0.9286	1	0.9286
3	0.9286	0.9286	0.9286
4	1	0.9286	1
5	1	0.9286	1
6	0.9286	0.9286	0.9286
7	1	1	1
8	0.9286	0.9286	0.9286
9	1	0.9286	1
10	0.9286	0.9286	0.9286
Average	0.9643	0.95	0.9643

Table 5.5 demonstrated the accuracy of SVM classifier for each gene expression dataset with multiple runs. The gene expression dataset that involved genes in all biclusters and the original gene expression achieved 96.43 percent accuracy respectively. Meanwhile, the gene expression dataset that involved the genes that occurred in more than one bicluster achieved 95 percent accuracy. The results obtained highlighted that the model's performance was affected by the different split of data for training and testing. Furthermore, the variation of the accuracy score across multiple

runs demonstrated that the performance was not excessively dependent on the random split of data. Instead, the variation can be due to the diversity in bicluster pattern. As the genes from different biclusters were combined into a dataset for classification, the diversity of biological patterns can influence the model's ability to generalize the unseen data across different genes' patterns as different biclusters showed distinct pattern of genes to EC.

Table 5.6: Performance Measurement Based on Confusion Matrix

Metrics	Gene Expression Dataset That Involved Genes		
	In All Bicluster	Occur In More Than One Bicluster	Original Dataset
No. of Features	283 genes	3 genes	2735 genes
Accuracy	0.9643	0.95	0.9643
Precision	1	0.975	1
Recall	0.9286	0.9286	0.9286
Specificity	1	0.9712	1
F1 Score	0.9616	0.9482	0.9616

Table 5.6 represented the evaluation of the SVM classifier for three different gene expression datasets. Firstly, the gene expression dataset that involved genes in all bicluster achieved 96.43 percent accuracy and the F1 score value at 96.16 percent with 283 genes. Secondly, the gene expression dataset that involved genes that occurred in more than one bicluster achieved 95 percent of the accuracy and reached the F1 score value at 0.9482 with 3 genes. Thirdly, the original gene expression dataset showed 96.43 percent accuracy and 0.9616 F1 score with 2735 genes. In conclusion, the performance of SVM classifier indicated that biclustering algorithm further improved the gene selection process by selecting the important features, thereby increasing the performance of classification as the features for gene expression dataset that involved genes in all biclusters and gene expression dataset that involved genes that occurred in more than one bicluster are less than the original gene expression dataset.

5.5 Gene Validation

For your information, the goal of plaid biclustering algorithm is to discover the key features by showing the patterns in gene and samples data. Hence, the genes found inside the bicluster were important indicators that predictive of EC. Furthermore, the gene expression dataset involved genes that occurred in more than one bicluster was chosen as the better dataset based on the findings of performance measurement by confusion matrix. The dataset not only exhibit the good performance compared to others, but also the frequently involvement in multiple biclusters indicated that these genes consistently aligned with the biological patterns found in the data. This situation made the genes a significant choice for gene validation. The genes found in the gene expression dataset involved genes that occurred in more than one bicluster were EPHB4, LAMB3 and HOXD11.

5.5.1 EPHB4

According to Hasina et al (2013) on the study of immunohistochemistry, EPHB4 is considerably overexpressed in esophageal cancer cell lines and primary tumor tissues which demonstrates that its levels are raised in disease states. Furthermore, an increase of EPHB4 gene copy numbers in some esophageal cancer samples and cell lines suggests a genetic foundation for its higher level of expression in tumors (Hasina et al, 2013). Although there are no significant changes in EPHB4 within cancer cells, its functional overexpression and activity are influenced by other focused oncogenic drivers, emphasizing its significance as an inhibitor of treatment for esophageal cancer (Hasina et al, 2013). The disruption of EPHB4's standard control of expression underscores the protein's significance as a biomarker for disease diagnosis and targeted therapy, as well as its possible role in the biology of esophageal cancer (Hasina et al, 2013).

Furthermore, a study on exploring the roles of cation-dependent mannose 6-phosphate receptor (M6PR) and ephrin B type receptor 4 (EphB4) in serine (SRGN) exosomes in promoting tumor angiogenesis and invasion of ESCC cells had been

carried out by Yan et al (2023). Based on the findings, exosomes generated from ESCC cells that overexpressed SRGN showed higher amounts of EPHB4, indicating a potential role for this protein in the development of cancer (Yan et al, 2023). Significantly, exosome EPHB4 increased ESCC cells' capacity for invasion, indicating a potential function in tumor malignancy and metastasis (Yan et al, 2023). Furthermore, the significant association between EphB4 expression and SRGN levels in ESCC patients' serum highlights its potential as a prognostic indicator, with high serum EphB4 being associated with lower overall survival (Yan et al, 2023).

5.5.2 LAMB3

A study on the assessing the expression of LAMB3 in esophageal cancer stem cell and adherent cells had been done by Ehtesham et al (2022). The study suggested that the involvement of LAMB3 in the development of esophageal cancer stem cells (CSCs) and the advancement of tumours highlights its significance as a potential cause of the cancer (Ehtesham et al, 2022). The different expression pattern of CSCs and adherent cells shows that it is involved in critical processes such as spheroid formation, CSC development, and tumour growth (Ehtesham et al, 2022). Hence, LAMB3 might be a good target for treatments which use to prevent CSC-driven tumour growth and spreading in esophageal cancer (Ehtesham et al, 2022).

The research also explained that LAMB3 helps to produce Laminin-332, an important extracellular matrix (ECM) protein for the CSC microenvironment (Ehtesham et al, 2022). Downregulation of LAMB3 in esophageal CSCs has been linked to increased sphere formation, implying a role in enhancing CSC traits such as self-renewal and tumorigenicity (Ehtesham et al, 2022). Laminin-332, which includes LAMB3, has been linked to cancer invasion, migration, and metastasis which possibly are one of the factors that gave rise to EC cancer. (Ehtesham et al, 2022).

5.5.3 HOXD11

According to the National Library of Medicine (2024), HOXD11 is belongs to the homeobox (HOX) gene family, which consists of transcription factors that are important for morphogenesis in a variety of organisms with multiple cells and are highly conserved. The HOX gene family is important for embryonic development, and its dysregulation is associated with several malignancies, including esophageal cancer (Akbar, Zhang and Liu, 2023). When HOX genes are dysregulated, their normal developmental functions are disrupted, causing cancer cells to behave abnormally (Akbar et al, 2023). Dysregulated HOX genes, such as HOXD11, may affect cell proliferation, metastasis, and treatment resistance of cancer cells, thereby affecting tumor progression and patient prognosis (Akbar et al, 2023).

Dysregulation of HOX genes, such as HOXD11, can have a significant impact on cancer biology (Akbar et al, 2023). The overexpression of HOX genes can lead to uncontrolled growth of cells, which can enable tumors to spread quickly and escape regulatory systems that typically prevent excessive cell division (Akbar et al, 2023). Dysregulated HOX genes can also help cancer cells spread to distant regions of the body and improve their capacity for metastasis (Akbar et al, 2023). As a result, misregulation of HOX genes enhances the complexity of cancer development and creates major difficulties for the treatment (Akbar et al, 2023).

5.6 Additional Testing

5.6.1 Investigating SVM Classification Effectiveness on Each Bicluster

In this study, the experiments focused on filtering the gene expression dataset by combining the genes found in the bicluster together. However, a further analysis had been conducted to investigate the effectiveness of SVM on each bicluster. The procedure started by filtering the data from the gene expression dataset by the genes in each bicluster. Four gene expression datasets were analyzed using SVM classifier to evaluate the performance of each bicluster by ten-fold cross validation and

confusion matrix. This step is to understand the effectiveness of SVM to learn the specific gene expression patterns to classify the data.

The results presented in Table 5.7 indicated an improvement in performance compared to the study. Bicluster 1 achieved an accuracy of 95 percent and 94.49 percent of F1 score. Besides that, bicluster 2 demonstrated the strongest performance among four biclusters with 98.57 percent of accuracy and 98.46 percent of F1 score. Additionally, bicluster 3 indicated accuracy at 0.90 with 0.8906 F1 score. Lastly, bicluster 4 achieved 95.72 percent of accuract and 0.9539 of F1 score. In general, the performance was better than the study due to the bicluster focused on the gene expression patterns that exhibit common behavior under specific conditions. This expression patterns of bicluster provided a clearer insight for SVM classifier to learn from the data and resulted in enhancing the ability to classify the samples accurately. In contrast, the gene expression dataset contained more diverse expression patterns.

Table 5.7: Performance Measurement Based on Multiple Run and Confusion Matrix

	Gene Expression Dataset That Involved Genes In			
	Bicluster 1	Bicluster 2	Bicluster 3	Bicluster 4
Multiple Run of SVM Classifier				
Accuracy	0.95	0.9857	0.9	0.9572
Confusion Matrix				
Accuracy	0.95	0.9857	0.9	0.9572
Precision	1	1	0.9732	1
Recall	0.9	0.9714	0.8286	0.9143
Specificity	1	1	0.9714	0.9857
F1 Score	0.9449	0.9846	0.8906	0.9539

5.6.2 Applying the Experiment to New Gene Expression Dataset

A new gene expression dataset that indicated the expression pattern of ovarian cancer had been retrieved. After the data preprocessing and gene selection process, the dimension dataset become 67 samples with 2140 genes. The varying scales in the data suggested the presence of noise within the ovarian cancer dataset. The inconsistency can be due to the biological differences among the samples. As the genes can be detected in various tissues, body fluids and blood, hence their activities can be varied depending on the specific conditions within samples.

	Gene Symbol	GSM94332	GSM94339	GSM94341	GSM94344	GSM94346	GSM94347	GSM94348	GSM94349
0	HSPA6	142.7	67.0	27.1	87.5	188.0	262.0	171.4	99.6
1	GUCA1A	33.6	61.7	30.3	32.6	39.5	35.4	57.8	47.3
2	MMP14	662.6	483.5	266.6	697.9	850.7	734.4	393.4	1571.7
3	TRADD	316.9	300.8	391.9	1055.9	398.6	389.6	416.6	369.5
4	PLD1	130.2	130.6	115.0	89.5	67.3	73.0	254.1	97.9

Figure 5.7: Ovarian Cancer Dataset after Gene Selection Process

Applying normalization to the ovarian cancer dataset is essential to ease the effects of noise and different scaling of data. By standardizing the gene expression data able to improve the reliability of the biclustering and classification method. This is because normalization ensured the data was transformed into a comparable scale and reduced the bias among samples. Figure below showed the ovarian cancer dataset after normalization.

	Gene Symbol	GSM94332	GSM94339	GSM94341	GSM94344	GSM94346	GSM94347	GSM94348
0	HSPA6	0.004800	0.002011	0.001385	0.003290	0.006445	0.009737	0.006028
1	GUCA1A	0.001107	0.001850	0.001552	0.001209	0.001332	0.001303	0.002014
2	MMP14	0.022398	0.014663	0.013946	0.026424	0.029260	0.027321	0.013871
3	TRADD	0.010697	0.009113	0.020517	0.039993	0.013695	0.014487	0.014691
4	PLD1	0.004377	0.003943	0.005995	0.003366	0.002289	0.002702	0.008949

Figure 5.8: Ovarian Cancer Dataset after Normalization

Based on the result in Table 5.8 indicated that the involvement of PPI data significantly improved the accuracy of the ovarian cancer dataset analysis. Furthermore, table 5.9 demonstrated that using a test size of 0.1 test size generated the better result for the SVM classifier. This suggests that a smaller test set which only comprising 10 percent of the data allowed for more effective training and validation.

This is because the model able to learn from larger number of data which lead to a better generalization.

Table 5.8: Accuracy of Ovarian Cancer Dataset with and Without PPI Data

Test Size	Ovarian Cancer Dataset	
	Without PPI Data	With PPI Data
0.1	0.8571	0.8571
0.2	0.7857	0.7857
0.3	0.8095	0.8095
0.4	0.7407	0.8148
Average	0.7983	0.8168

Table 5.9: Accuracy of Ovarian Cancer Dataset with Different Random State and Test Size

Run	Accuracy with Different Test Size			
	0.1	0.2	0.3	0.4
1	0.8571	0.8571	0.8095	0.8148
2	1	0.7143	0.7143	0.6667
3	1	0.8571	0.8571	0.7037
4	0.7143	0.7143	0.6667	0.7407
5	0.7143	0.8571	0.7143	0.7407
6	0.8571	0.7857	0.8095	0.8148
7	0.8571	0.8571	0.8095	0.7778
8	0.7143	0.8571	0.8571	0.8519
9	0.8571	0.7857	0.8095	0.8148
10	0.7143	0.7857	0.7619	0.8148
Average	0.8428	0.8071	0.7809	0.7741

After applying the Plaid Biclustering Model to the ovarian cancer dataset, three biclusters were generated. The genes identified within the biclusters were then used to

create two distinct datasets for further analysis. Figures below showed the biclusters formed by Plaid Biclustering Model.

Gene Symbol	BACH2	PRMT2	KCNMA1	ADAMTS1	CHST3	PRELP	PRDX2	Target
GSM94332	0.067106	0.092070	0.096593	0.283825	0.080160	0.182937	0.574880	0
GSM94339	0.044100	0.166408	0.100546	0.353890	0.279701	0.131840	0.412001	0
GSM94341	0.024932	0.185145	0.021065	0.026433	0.132625	0.082299	0.579574	0
GSM94344	0.057520	0.216188	0.258553	0.278014	0.259472	0.177095	0.576480	0
GSM94346	0.049674	0.163660	0.277762	0.304385	0.210223	0.339423	0.578648	0

Figure 5.9: Ovarian Cancer Dataset - Bicluster 1

Gene Symbol	MUC1	KRT8	SLC35A2	MTX1	PSMD14	MTCH2	RIPK4	Target
GSM94344	0.576480	0.576480	0.417910	0.709127	0.432721	0.502354	0.780367	0
GSM94346	0.578648	0.578648	0.269406	0.533280	0.578648	0.618510	0.275802	0
GSM94350	0.097479	0.062923	0.129398	0.365731	0.328309	0.341490	0.019542	1
GSM94352	0.577740	0.872684	0.247037	0.323553	0.723104	0.790185	0.240996	0
GSM94374	0.127823	0.104086	0.134605	0.426404	0.258578	0.285476	0.032071	1

Figure 5.10: Ovarian Cancer Dataset - Bicluster 2

Gene Symbol	FANCA	Target
GSM94352	0.138362	0
GSM94395	0.097620	0
GSM94411	0.134773	0

Figure 5.11: Ovarian Cancer Dataset - Bicluster 3

Table 5.11 indicated the datasets involving genes in all biclusters demonstrated the accuracy value at 0.8071. In contrast, the original ovarian cancer dataset showed a slightly lower accuracy of 79.29 percent indicating that biclustering algorithm provide an improvement in classification performance. From the result, a conclusion that ovarian cancer dataset that involved genes in all bicluster achieved the better result in the classification performance.

Table 5.10: The Dimension of Three Different Ovarian Cancer Dataset

Ovarian Cancer Dataset	Samples	Genes
Genes In All Bicluster	67	129
Original Dataset	67	2140

Table 5.11: Cross Validation of Different Ovarian Cancer Dataset

	Ovarian Cancer Dataset That Involved Genes	
	In All Bicluster	Original Dataset
Fold 1	1	1
Fold 2	0.8571	0.8571
Fold 3	0.7143	0.5714
Fold 4	0.7143	0.8571
Fold 5	0.5714	0.7143
Fold 6	0.8571	0.5714
Fold 7	0.8571	0.8571
Fold 8	0.8333	0.8333
Fold 9	0.8333	0.6667
Fold 10	0.8333	1
Average	0.8071	0.7929

The datasets were further analysed by multiple runs to evaluate the consistency of the SVM performance. This approach to ensure obtained with a more reliable result of the model's effectiveness. Table 5.12 indicated the accuracy of two different ovarian dataset across multiple runs. The result demonstrated that the ovarian cancer dataset that involved genes in all bicluster achieved an accuracy of 75.71 percent while the original ovarian cancer dataset achieved 79.29 percent accuracy. By considering the dimensions of two datasets, the dataset with genes in all bicluster consisted of 67 samples with 129 genes with a slightly lower accuracy than the original dataset showed that the genes selected through the biclusters contributed to the effectiveness of SVM. The gene selection process enabled the classifier to focus on the important indicators across multiple samples. The model able to have a clearer understanding of the genes' pattern in the dataset with genes in all bicluster rather than the original dataset.

Table 5.12: Accuracy of Three Different Dataset with Different Random State

Run	Ovarian Cancer Dataset That Involved Genes	
	In All Bicluster	Original Dataset
1	0.7143	1
2	0.5714	0.8571
3	1	0.5714
4	0.7143	0.8571
5	0.7143	0.7143
6	0.8571	0.5714
7	0.8571	0.8571
8	1	0.8333
9	0.5714	0.6667
10	0.5714	1
Average	0.7571	0.7929

Table 5.13 provided the information on the classification performance of two distinct ovarian cancer dataset. However, the concerning aspect raised from the precision, recall and F1 Score as the dataset a lower score. These metrics suggest a significant issue with the model's performance as it failed to correctly identify any positive samples. The presence of noisy data, outliers and different scaling across the ovarian cancer dataset misrepresent the model's ability to learn the meaningful patterns. Secondly, the imbalanced class distribution within the dataset with only 13 out of 67 samples representing the normal class. The imbalanced class distribution bias the model towards the majority class which is tumour and resulting in poor performance on the normal tissues. Hence, to address these challenges, preprocessing techniques such as outlier detection and replacement, noise reduction and balancing the class distribution must be carried out to derive more accurate insights for effective treatment and disease diagnosis. However, despite these challenges, the biclustering algorithm had demonstrated its ability to enhance the performance in identifying potential biomarkers.

Table 5.13: Confusion Matrix of Two Different Dataset

Metrics	Ovarian Cancer Dataset that Involved Genes	
	In All Bicluster	Original Dataset
Accuracy	0.7571	0.7714
Precision	0.2	0.2
Recall	0.2	0.2
Specificity	0.85	0.8667
F1 Score	0.2	0.2

5.7 Chapter Summary

In conclusion, the involvement of PPI data with gene expression datasets significantly enhances the accuracy of identifying potential biomarkers for EC as the gene selection process filters out the insignificant features. Furthermore, using 60 percent of gene expression dataset for training provides enough data to train the model and capture underlying patterns. Additionally, plaid biclustering algorithm grouped related genes making it efficient in highlighting the potential biomarkers. However, it is important to preprocess the dataset before training the model to ensure the optimal results.

CHAPTER 6

CONCLUSION & RECOMMENDATION

6.1 Introduction

The research mainly focused on the identification of potential biomarkers of EC using plaid biclustering algorithm. The experiment had been carried out by using gene expression dataset with PPI data. Plaid biclustering model form a collection of biclusters with the underlying patterns of the data. The performance of each biclusters had been further observed and verified by SVM classifier. The results showed that biclustering algorithm able to classify the genes with the similar features and able to identify the biomarker for EC.

6.2 Achievements

The achievements for this research are generate the input data from gene expression and PPI data, implement the biclustering algorithm in identification of potential biomarkers from the input data, evaluate the selected potential biomarkers using SVM classifier and verify the identified potential biomarkers with biological knowledge bases.

6.2.1 Objective 1: To derive input data from gene expression and PPI data

In this step, data preprocessing had been carried out to remove the missing genes and obtained the average value of duplicated genes. Then, gene selection process was applied to the gene expression dataset in order to filter out the insignificant features. The filtering process of gene expression dataset was based on the genes in

PPI data. After the gene selection process, the dimensions of the gene expression dataset become 2735 genes with 34 samples.

6.2.2 Objective 2: To implement biclustering algorithms in identification of potential biomarkers from the derived input data

The optimum number of biclusters was identified by using the elbow method. The Elbow method is the technique based on KMeans algorithm to retrieve the sum of square error with the different number of clusters. The result of elbow method showed that the optimum number of biclusters for gene expression dataset is four. Then, a plaid biclustering model was implemented to form four biclusters. The general process of plaid biclustering model was found the coherent groups of genes and samples and captured the shared behaviour in a common layer. After that, the behaviour had been compared and verified before formed a bicluster. There are total of four biclusters were formed.

6.2.3 Objective 3: To evaluate the selected potential biomarkers using SVM through ten-fold cross validation and confusion matrix

The biclustering result indicated that the retrieved samples are cancerous. SVM classifier cannot classify the dataset when only consists of one target case. Since the goal of biclustering is to discover a group of genes and samples with similar features, hence the assumption that the genes found in the biclusters are important indicators of EC can be made. Hence, the genes in the biclusters were used to do another gene selection process. In this process, two gene expression datasets were formed which are gene expression dataset that involved genes in all biclusters and gene expression dataset that involved genes that occurred in more than one bicluster. Then, these two datasets were compared with the original gene expression dataset. The SVM classifier demonstrated with the result of gene expression dataset that involved genes that occurred in more than one bicluster is better than others.

6.2.4 Objective 4: To verify the identified potential biomarkers with biological knowledgebases such as NCBI

The genes found in the gene expression dataset that involved genes that occurred in more than one bicluster were further validated with the NCBI. The genes found inside the dataset are EPHB4, LAMB3 and HOXD11. These three genes showed their effort in the EC cell development after validated with NCBI.

6.3 Suggestion for Improvement and Future Works

There are still available improvements and future works can be done in this research. These includes:

(a) Development a method of determining the optimal pruning threshold value to be used in Plaid Model Biclustering Algorithm.

(b) Integration of machine learning techniques to enhance the performance and scalability of biclustering algorithms in handling high dimensional dataset.

(c) Perform hyperparameter tuning to enhance the model's performance.

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.

Akbar, A., Zhang, L. and Liu, H.S., (2023). 'Unlocking Esophageal Carcinoma's Secrets: An integrated Omics Approach Unveils DNA Methylation as a pivotal Early Detection Biomarker with Clinical Implications.' *medRxiv*, pp.2023-09.

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.

Bartha, Á. and Győrffy, B.. (2021) 'TNMplot. com: a web tool for the comparison of gene expression in normal, tumor and metastatic tissues.' *International journal of molecular sciences*, *22*(5), p.2622.

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.

Cherradi, B., Terrada, O., Ouhmida, A., Hamida, S., Raihani, A., & Bouattane, O. (2021, July). 'Computer-aided diagnosis system for early prediction of atherosclerosis using machine learning and K-fold cross-validation.' In *2021 international congress of advanced technology and engineering (ICOTEN)* (pp. 1-9). IEEE.

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.

Ehtesham, A., Khosravi, A., Jazi, M.S., Asadi, J. and Jafari, S.M., (2022). 'Decreased Expression of LAMB3 Is Associated with Esophageal Cancer Stem Cell Formation.' *Advanced Pharmaceutical Bulletin*, *12*(4), p.828.

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.

Fuadah, Y. N., Ubaidullah, I. D., Ibrahim, N., Taliningsing, F. F., Sy, N. K., & Pramuditho, M. A. (2022). 'Optimasi Convolutional Neural Network dan K-Fold Cross Validation pada Sistem Klasifikasi Glaukoma.' *ELKOMIKA: Jurnal Teknik Energi Elektrik, Teknik Telekomunikasi, & Teknik Elektronika, 10* (3), 728.

Hasina, R., Mollberg, N., Kawada, I., Mutreja, K., Kanade, G., Yala, S., Surati, M., Liu, R., Li, X., Zhou, Y., Ferguson, B. D., Nallasura, V., Cohen, K. S., Hyjek, E., Mueller, J., Kanteti, R., El Hashani, E., Kane, D., Shimada, Y., Lingen, M. W., … Salgia, R. (2013). 'Critical role for the receptor tyrosine kinase EPHB4 in esophageal cancers'. *Cancer research, 73*(1), 184–194. doi: 10.1158/0008-5472.CAN-12-0915

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.

Keerthana, D., Venugopal, V., Nath, M.K. and Mishra, M.. (2023) 'Hybrid convolutional neural networks with SVM classifier for classification of skin cancer.' *Biomedical Engineering Advances, 5*, p.100069.

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).

National Library of Medicine. (2024). *HOXD11 homeobox D11 [Homo sapiens (human)]*. National Center for Biotechnology Information. Available at: https://www.ncbi.nlm.nih.gov/gene/3237 (Accessed: 3 May 2024).

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.

Ozer, M.E., Sarica, P.O. and Arga, K.Y.. (2020) 'New machine learning applications to accelerate personalized medicine in breast cancer: rise of the support vector machines.' *Omics: a journal of integrative biology*, *24*(5), pp.241-246.

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., Wãs, 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.

Scikit Learn. (no date). *sklearn.metrics.f1_score*. Available at: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html (Accessed: 26 April 2024)

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.

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE access*, *8*, pp. 80716-80727.

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.

Vujović, Ž.. (2021) 'Classification model evaluation metrics.' *International Journal of Advanced Computer Science and Applications*, *12*(6), pp.599-606.

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.

Yan, D., Cui, D., Zhu, Y., Chan, C.K.W., Choi, C.H.J., Liu, T., Lee, N.P., Law, S., Tsao, S.W., Ma, S. and Cheung, A.L.M., (2023). 'M6PR-and EphB4-rich exosomes secreted by serglycin-overexpressing esophageal cancer cells promote cancer progression.' *International Journal of Biological Sciences*, *19*(2), p.625.

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.

Appendix A Figures of the Experiment's Output

Github link to the code: https://github.com/wyu04/FYP

A - Gene In All Bicluster
B - Gene In Multiple Bicluster
C - Original Dataset

Random State		5	10	15	20	25	30	35	40	45	50
Accuracy	A	1	0.9286	0.9286	1	1	0.9286	1	0.9286	1	0.9286
	B	1	1	0.9286	0.9286	0.9286	0.9286	1	0.9286	0.9286	0.9286
	C	1	0.9286	0.9286	1	1	0.9286	1	0.9286	1	0.9286
Precision	A	1	1	1	1	1	1	1	1	1	1
	B	1	1	1	0.875	0.875	1	1	1	1	1
	C	1	1	1	1	1	1	1	1	1	1
Recall	A	1	0.8571	0.8571	1	1	0.8571	1	0.8571	1	0.8571
	B	1	1	0.8571	1	1	0.8571	1	0.8571	0.8571	0.8571
	C	1	0.8571	0.8571	1	1	0.8571	1	0.8571	1	0.8571
Specificity	A	1	1	1	1	1	1	1	1	1	1
	B	1	1	1	0.8571	0.8571	1	1	1	1	1
	C	1	1	1	1	1	1	1	1	1	1
F1 Score	A	1	0.9231	0.9231	1	1	0.9231	1	0.9231	1	0.9231
	B	1	1	0.9231	0.9333	0.9333	0.9231	1	0.9231	0.9231	0.9231
	C	1	0.9231	0.9231	1	1	0.9231	1	0.9231	1	0.9231

Figure 1: Confusion Matrix Result of EC Cancer Dataset with Different Random State

A - Bicluster 1
B - Bicluster 2
C - Bicluster 3
D - Bicluster 4

Random State		5	10	15	20	25	30	35	40	45	50
Accuracy	A	1	0.9286	0.9286	1	0.9286	0.8571	1	0.9286	1	0.9286
	B	1	1	0.9286	1	1	1	1	0.9286	1	1
	C	0.9286	0.9286	0.9286	0.9286	0.8571	0.8571	0.9286	0.8571	0.9286	0.8571
	D	1	0.9286	0.9286	1	1	0.9286	1	0.9286	0.9286	0.9286
Precision	A	1	1	1	1	1	1	1	1	1	1
	B	1	1	1	1	1	1	1	1	1	1
	C	1	1	1	0.875	0.8571	1	1	1	1	1
	D	1	1	1	1	1	1	1	1	1	1
Recall	A	1	0.8571	0.8571	1	0.8571	0.7143	1	0.8571	1	0.8571
	B	1	1	0.8571	1	1	1	1	0.8571	1	1
	C	0.8571	0.8571	0.8571	1	0.8571	0.7143	0.8571	0.7143	0.8571	0.7143
	D	1	0.8571	0.8571	1	1	0.8571	1	0.8571	0.8571	0.8571
Specificity	A	1	1	1	1	1	1	1	1	1	1
	B	1	1	1	1	1	1	1	1	1	1
	C	1	1	1	0.8571	0.8571	1	1	1	1	1
	D	1	0.8571	1	1	1	1	1	1	1	1
F1 Score	A	1	0.9231	0.9231	1	0.9231	0.8333	1	0.9231	1	0.9231
	B	1	1	0.9231	1	1	1	1	0.9231	1	1
	C	0.9231	0.9231	0.9231	0.9333	0.8571	0.8333	0.9231	0.8333	0.9231	0.8333
	D	1	0.9231	0.9231	1	1	0.9231	1	0.9231	0.9231	0.9231

Figure 2: Confusion Matrix Result of EC Cancer Bicluster Dataset with Different Random State

A - Gene In All Bicluster
B - Gene In Multiple Bicluster
C - Original Dataset"

Random State		5	10	15	20	25	30	35	40	45	50
Accuracy	A	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571
	B	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571
	C	0.7143	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571
Precision	A	0	0	0	0	0	0	0	0	0	0
	B	0	0	0	0	0	0	0	0	0	0
	C	0	0	0	0	0	0	0	0	0	0
Recall	A	0	0	0	0	0	0	0	0	0	0
	B	0	0	0	0	0	0	0	0	0	0
	C	0	0	0	0	0	0	0	0	0	0
Specificity	A	1	1	1	1	1	1	1	1	1	1
	B	1	1	1	1	1	1	1	1	1	1
	C	0.8333	1	1	1	1	1	1	1	1	1
F1 Score	A	0	0	0	0	0	0	0	0	0	0
	B	0	0	0	0	0	0	0	0	0	0
	C	0	0	0	0	0	0	0	0	0	0

Figure 3: Confusion Matrix of Ovarian Cancer Dataset with Different Random State

A - Bicluster 1
B - Bicluster 2
C - Bicluster 3
D - Bicluster 4

Random State		5	10	15	20	25	30	35	40	45	50
Accuracy	A	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571
	B	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571
	C	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571
	D	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571	0.8571
Precision	A	0	0	0	0	0	0	0	0	0	0
	B	0	0	0	0	0	0	0	0	0	0
	C	0	0	0	0	0	0	0	0	0	0
	D	0	0	0	0	0	0	0	0	0	0
Recall	A	0	0	0	0	0	0	0	0	0	0
	B	0	0	0	0	0	0	0	0	0	0
	C	0	0	0	0	0	0	0	0	0	0
	D	0	0	0	0	0	0	0	0	0	0
Specificity	A	1	1	1	1	1	1	1	1	1	1
	B	1	1	1	1	1	1	1	1	1	1
	C	1	1	1	1	1	1	1	1	1	1
	D	1	1	1	1	1	1	1	1	1	1
F1 Score	A	0	0	0	0	0	0	0	0	0	0
	B	0	0	0	0	0	0	0	0	0	0
	C	0	0	0	0	0	0	0	0	0	0
	D	0	0	0	0	0	0	0	0	0	0

Figure 4: Confusion Matrix of Ovarian Cancer Bicluster Dataset with Different Random State

```
The number of outliers:  15184
Outlier values: [ 1866.7         4879.9         2504.2         ... 10484.4
 10176.76666667  9625.8        ]
The number of normal samples:  13
The dimension of ovarian cancer dataset:  (2140, 68)
```

Figure 5: The Noise and Class Imbalance Distribution of Ovarian Cancer Dataset

Random State	Accuracy with Different Test Size			
	0.1	0.2	0.3	0.4
5	0.8571	0.8571	0.8095	0.8148
10	1	0.7143	0.7143	0.6667
15	1	0.8571	0.8571	0.7037
20	0.7143	0.7143	0.6667	0.7407
25	0.7143	0.8571	0.7143	0.7407
30	0.8571	0.7857	0.8095	0.8148
35	0.8571	0.8571	0.8095	0.7778
40	0.7143	0.8571	0.8571	0.8519
45	0.8571	0.7857	0.8095	0.8148
50	0.8571	0.7857	0.7619	0.8148
Average	**0.84284**	0.80712	0.78094	0.77407

Figure 6: Accuracy of Ovarian Cancer Dataset with Different Test Size and Random State after Outlier Replacement

	Gene In All Bicluster	Original Dataset
Accuracy	0.7571	0.7714
Precision	0.2	0.2
Recall	0.2	0.2
Specificity	0.85	0.8667
F1 Score	0.2	0.2

Dataset	Gene In All Bicluster	Original Dataset
Sample	67	67
Genes	129	2140
Accuracy	75.71%	77.14%

Figure 7: Performance Measurement of Ovarian Cancer Dataset after Outlier Replacement

	Bicluster 1	Bicluster 2	Bicluster 3
Accuracy	0.8	0.8571	0.8571
Precision	0.225	0	0
Recall	0.3	0	0
Specificity	0.9167	1	1
F1 Score	0.24	0	0
Sample	67	67	67
Gene	103	25	1

Dataset	Bicluster 1	Bicluster 2	Bicluster 3
Accuracy	80.00%	85.71%	85.71%
Sample	67	67	67
Gene	103	25	1

Figure 8: Performance Measurement of Each Bicluster after Outlier Replacement