UTM
UNIVERSITI TEKNOLOGI MALAYSIA

FACULTY OF COMPUTING
UTM Johor Bahru

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

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

Presented by Gui Yu Xuan

Supervised by Dr. Chan Weng Howe

Presentation Link: https://youtu.be/Mo3of-sMN1Q

Demo Link: https://youtu.be/3dbpXcTMmt0

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Table Of Contents

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

INTRODUCTION

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Introduction

- Esophageal cancers
- The **combination of information on PPI and gene expression** enables the discovery of possible biomarkers and advances our understanding of disease.
- 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.
- **Biclustering method** can be used to analyse esophageal cancers through the gene expression data and PPI data **to detect gene clusters that exhibit differential expression** when compared to normal tissue in esophageal tumours.

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Problem Background

- 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.
- Biclustering can **decrease the high-dimensional character** of gene expression datasets by focusing on these co-expressed genes
- 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.**

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Problem Statement

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.

- Using only synthetic data to find biomarkers can produce false-positive results and overfit the data.
- Determine the biological significance of the data able to increase the possibility of discovering a true and informative biomarker.
- PPI and gene expression data are biological relevance data.

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

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.

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Research Objectives

- To generate input data from gene expression and protein-protein data.
- To implement biclustering algorithm in identification of potential biomarkers from the generated input data
- To evaluate the selected potential biomarkers using support vector machine through ten-fold cross validation and confusion matrix
- To verify the identified potential biomarkers with biological knowledgebases such as NCBI

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

LITERATURE REVIEW

Innovating Solutions

Biclustering Methods	Advantage	Disadvantage
• Correlated Pattern Biclustering (CPB)	• Work well in synthetic datasets • Perform well in large datasets	• Sensitive to noise • Sensitive in high differential expression
• QUBIC	• Better execution time	• Produce low acccurate and reliable result
• Bayesian Biclustering (BBC)	• Produce accurate and meaningful results even missing values presented	• Performance and execution time depends on noise level and sample size
• Binary inclusion-Maximal (BiMax)	• Effective for simple structure	• Performance based on sample size • Limited to discrete values datasets
• Plaid	• Capture Overlapped Bicluster • Low coherent variance	• Performance based on parameters

Biclustering Methods	Advantageous	Disadvantageous
• Iterative Signature Algorithm (ISA)	• Able to find homogenous group	• Give big impact on errors and outliers
• Spectral	• Able to identify unique molecular subtypes • Higher enrinchment analysis	• Performance based on noise level and sample size
• Order Preserving Submatrix (OPSM)	• Extract overlapped biclusters accurately • Produce stable output	• Do not filter output • Unable to analyse gene expression data adequately
• Cheng & Church (CC)	• Able to identify large number of biclusters	• Performance limited to higher noise level • Vulnerable to local optima • Long execution time

Classification Methods	Advantageous	Disadvantageous
• Support Vector Machine (SVM)	• Flexible • Handle High Dimensional Datasets	• Can Be Expensive and Complexity • Require Larger Computational Power
• K-Nearest Neighbours (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
• Neural Network	• can capture complex relationships. • flexible	• difficulty visualizing the decision-making process. • time consuming
• 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.

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Identifying Optimum Number of Clusters

- The quality metric for the calculation of number of clusters are **inertia** and **silhouette coefficient**.
- **Elbow method, silhouette method and gap statistic**
- Interpretation of elbow plots is sometimes subjective, the silhouette coefficient and gap statistical approaches can correctly quantify the number of clusters.
- **Gap statistics** include computations that could not always provide the same result.
- 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.

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

RESEARCH FRAMEWORK

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Phases

Phase 1

Research Planning and Initial Study

Phase 2

Development of Proposed Biclustering Methods

Phase 3

Evaluation of Potential Biomarkers by Classification Models

Phase 4

Verification of Selected Biomarkers with Biological Knowledgebases

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Gene Expression Dataset

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

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Protein Protein Interaction

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

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Phases

Phase 1

Research Planning and Initial Study

Phase 2

Development of Proposed Biclustering Methods

Phase 3

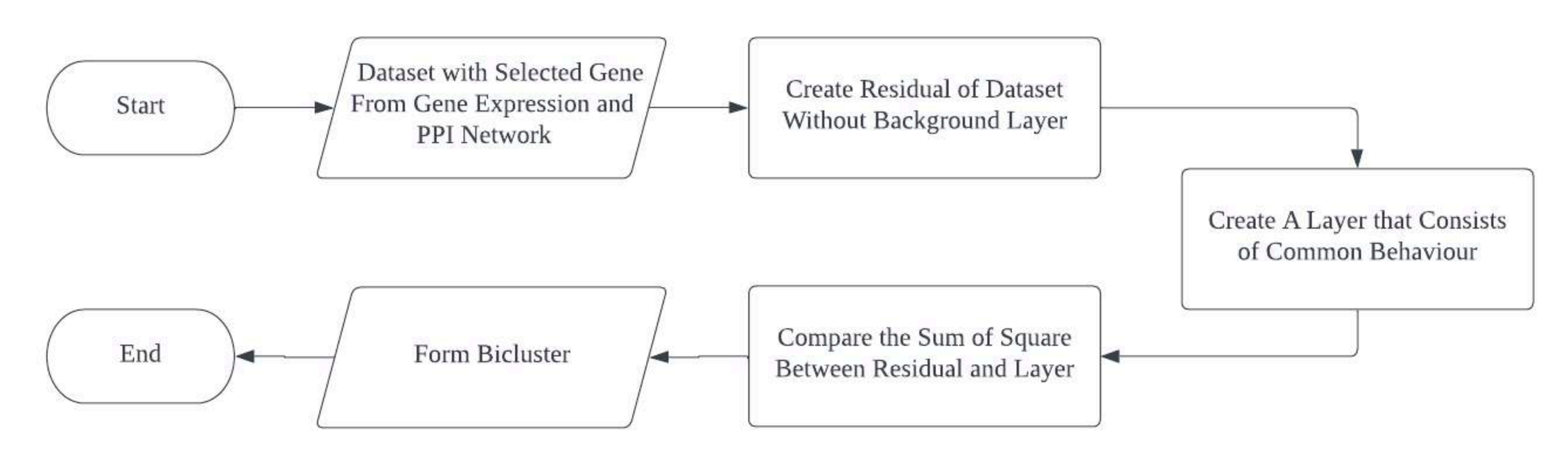
Evaluation of Potential Biomarkers by Classification Models

Phase 4

Verification of Selected Biomarkers with Biological Knowledgebases

Innovating Solutions

General Flow of Plaid Biclustering Model



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Phases

Phase 1

Research Planning and Initial Study

Phase 2

Development of Proposed Biclustering Methods

Phase 3

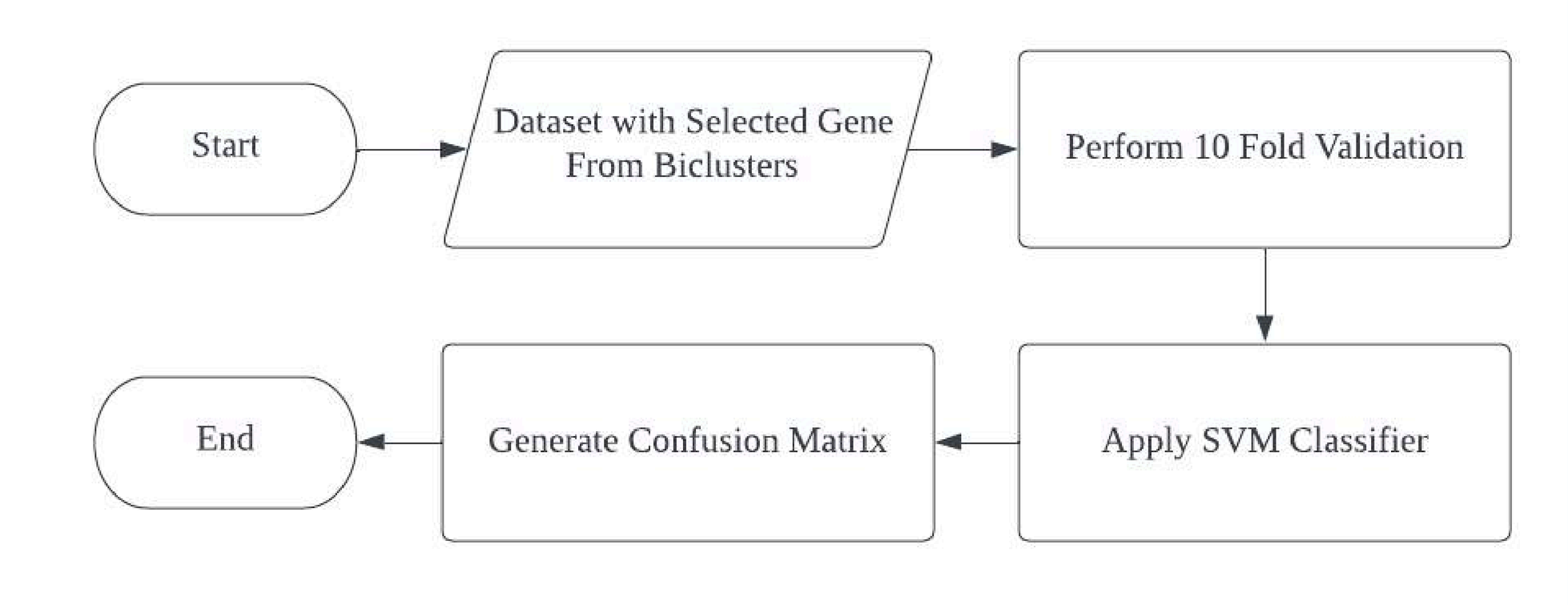
Evaluation of Potential Biomarkers by Classification Models

Phase 4

Verification of Selected Biomarkers with Biological Knowledgebases

Innovating Solutions

General Flow of Classification Method



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Phases

Phase 1

Research Planning and Initial Study

Phase 2

Development of Proposed Biclustering Methods

Phase 3

Evaluation of Potential Biomarkers by Classification Models

Phase 4

Verification of Selected Biomarkers with Biological Knowledgebases

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

RESEARCH DESIGN AND IMPLEMENTATION

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Development Process

The step to identify the potential biomarkers for EC cancer

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Data Preprocessing

Gene Expression Dataset

Gene	Sample 1	Sample 2	Sample 3
Gene 1	value	value	value
Gene 2	value	value	value
	value	value	value
Gene 3	value	value	value
Gene 1	value	value	value

Gene Expression Dataset

Gene	Sample 1	Sample 2	Sample 3
Gene 1	value	value	value
Gene 2	value	value	value
Gene 3	value	value	value
Gene 1	value	value	value

Gene Expression Dataset

Gene	Sample 1	Sample 2	Sample 3
Gene 1	value	value	value
Gene 2	value	value	value
Gene 3	value	value	value
Gene 1	value	value	value

Gene Expression Dataset

Gene	Sample 1	Sample 2	Sample 3
Gene 1	mean value	mean value	mean value
Gene 2	value	value	value
Gene 3	value	value	value

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Development Process

The step to identify the potential biomarkers for EC cancer

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Achieved Objective 1

To generate input data from gene expression and protein-protein interaction data

Gene Expression Dataset

Gene	Sample 1	Sample 2
Gene 1		
Gene 2		
Gene 3		
Gene 4		

PPI Data

Node 1	Node 2	Combined Score
Gene 1	Gene 2	
Gene 1	Gene 4	
Gene 2	Gene 1	
Gene 2	Gene 4	

Get the Gene List From PPI Data

Gene 1	Gene 2	Gene 4

Select The Data From Gene Expression Dataset Based On Gene List

Gene	Sample 1	Sample 2
Gene 1		
Gene 2		
Gene 4		

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

New Input After Gene Selection 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

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

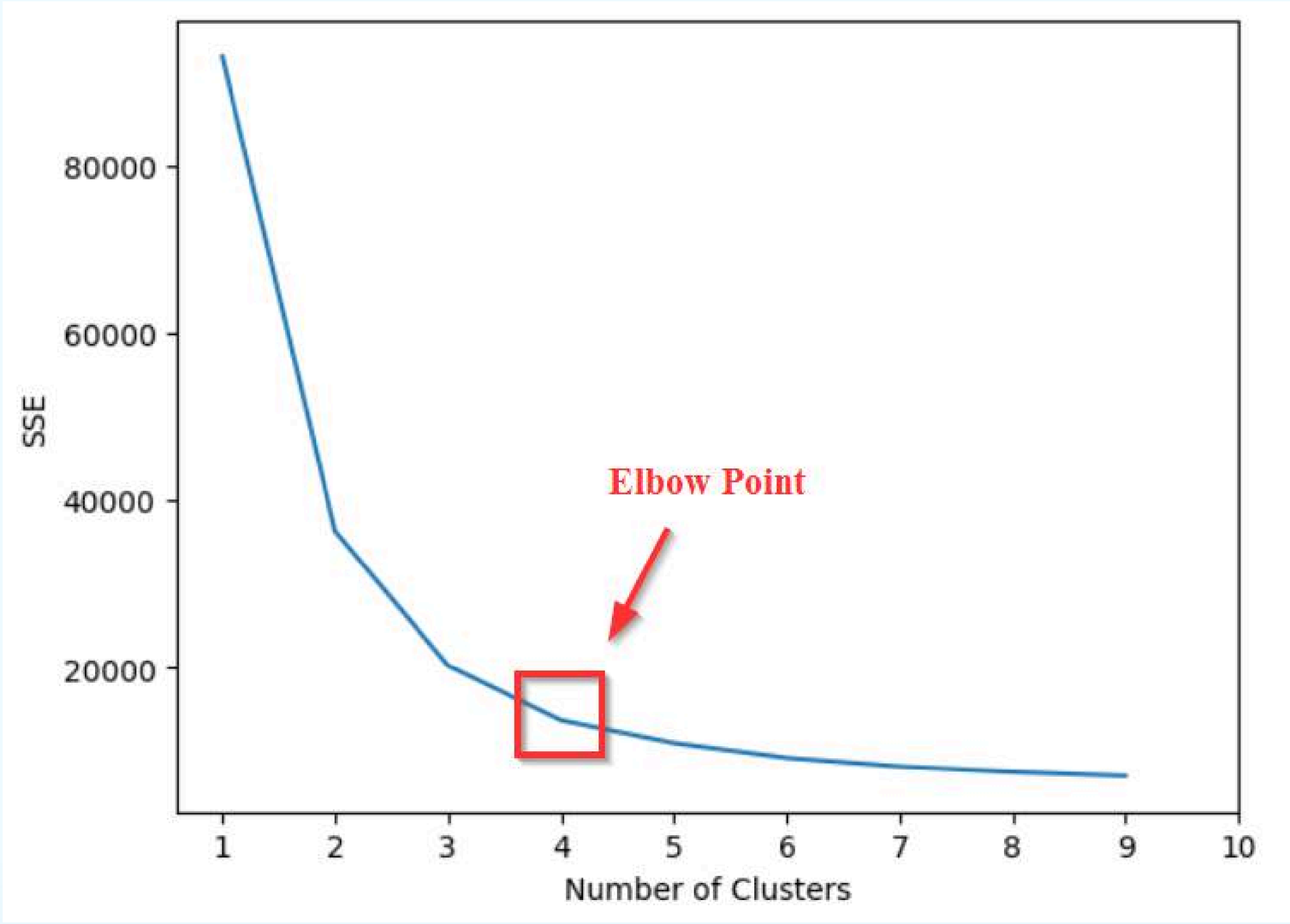
Development Process

The step to identify the potential biomarkers for EC cancer

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Purpose of Elbow Method

Identify the Optimum Number of Clusters To Be Used In Plaid Biclustering Model



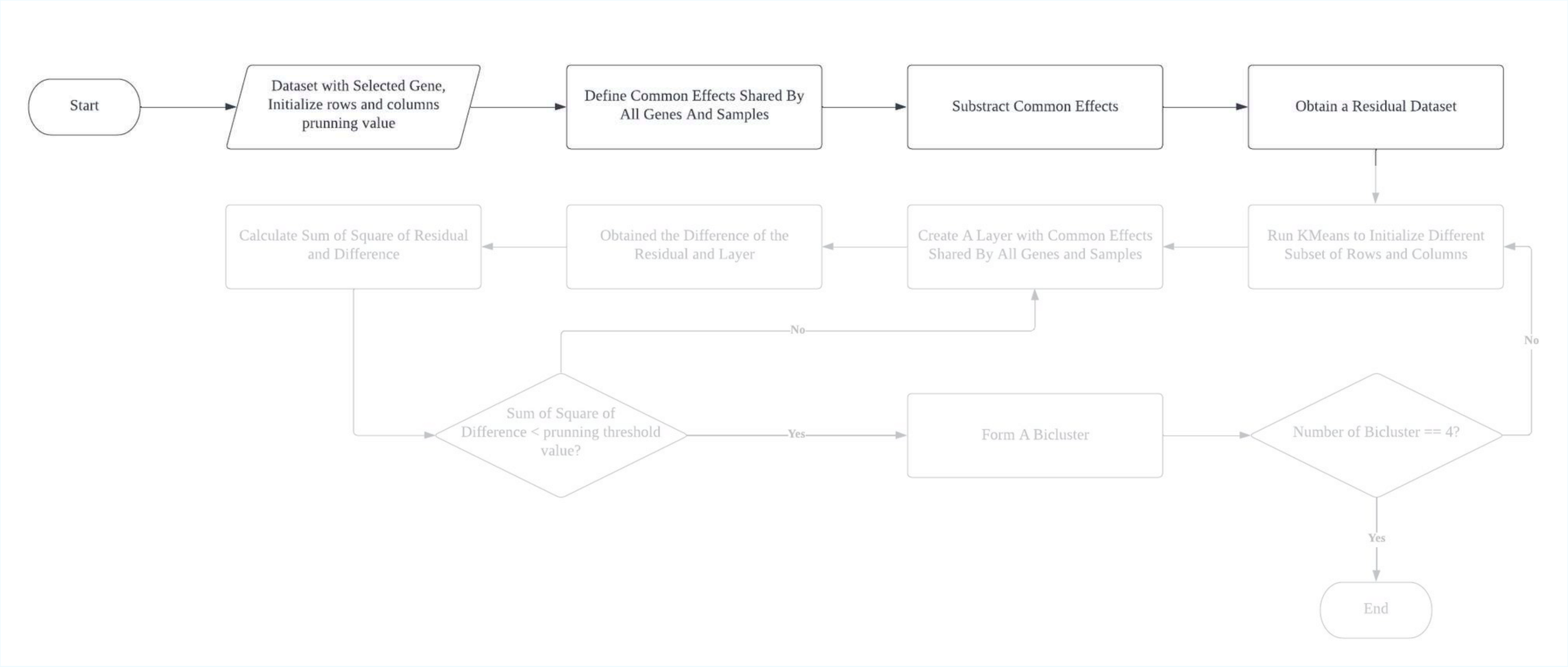
UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Development Process

The step to identify the potential biomarkers for EC cancer

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Plaid Model Biclustering Method



Innovating Solutions

Purpose of Common Effects Layer

Identify the Meaningful Data

Residual Dataset

Last Step: Using Gene Expression Dataset subtract Dataframe Hold Common Effects Value

Gene Expression Dataset

Gene	Sample 1	Sample 2	Sample 3
Gene 1	value	value	value
Gene 2	value	value	value
Gene 3	value	value	value
Gene 4	value	value	value

Define Common Effects Shared By All Genes And Samples

Gene	Sample 1	Sample 2	Sample 3	row_mean
Gene 1	value	value	value	row_mean 1
Gene 2	value	value	value	row_mean 2
Gene 3	value	value	value	row_mean 3
Gene 4	value	value	value	row_mean 4
col_mean	col_mean 1	col_mean 2	col_mean 3	overall mean

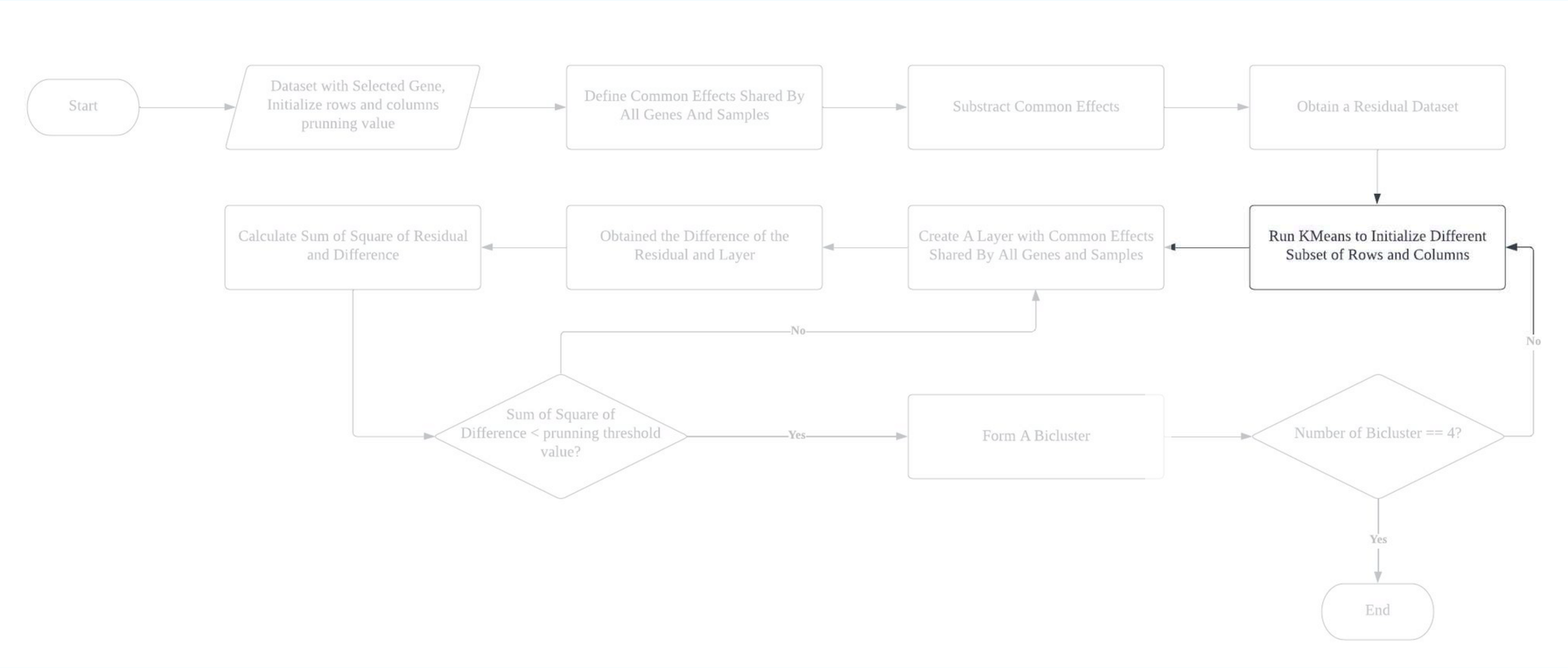
Notes: Overall Mean = Sum All Value/Number of Value

Create A Dataframe that Hold Common Effects Value

Gene	Sample 1	Sample 2
Gene 1	overall mean + row_mean 1 + col_mean 1	overall mean + row_mean 1 + col_mean 2
Gene 2	overall mean + row_mean 2 + col_mean 1	overall mean + row_mean 2 + col_mean 2

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Plaid Model Biclustering Method

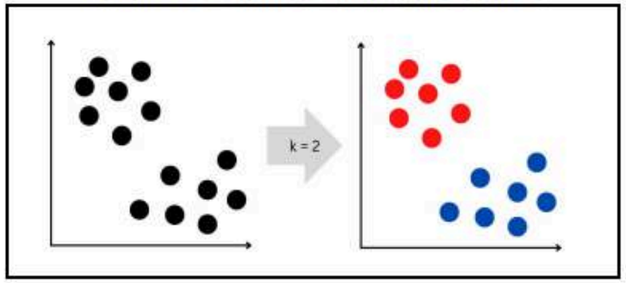


Innovating Solutions

Gene Expression Dataset

Gene	Sample 1	Sample 2	Sample 3
Gene 1	value	value	value
Gene 2	value	value	value
Gene 3	value	value	value
Gene 4	value	value	value

Perform KNN to Initialize Rows and Columns



Obtain Different Subset of Rows and Columns

Row 1	0	1	2
Column 1	0	2	3

Row 2	0	1
Column 3	3	

Gene Expression Dataset

Gene	Sample 1	Sample 2	Sample 3
Gene 1	value	value	value
Gene 2	value	value	value
Gene 3	value	value	value
Gene 4	value	value	value

Extract Out the Gene Expression Dataset

Gene	Sample 1	Sample 2	Sample 3
Gene 1	value	value	value
Gene 3	value	value	value
Gene 4	value	value	value

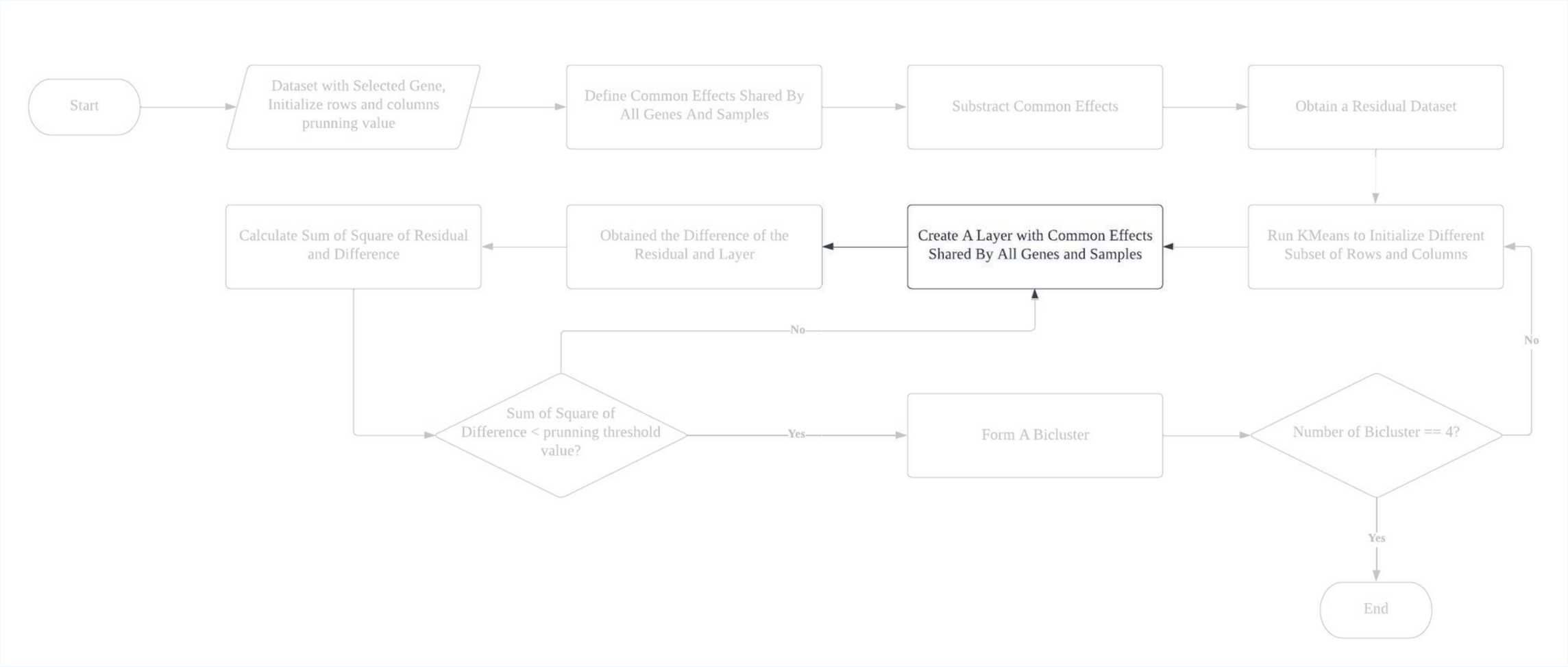
Gene	Sample 3
Gene 1	value
Gene 2	value

KMeans to Initialize Rows and Columns

Identify the Genes and Samples that Shared the Common Behaviour

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Plaid Model Biclustering Method



Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Purpose of Creating Common Effects Layer

For Comparing the Data Points With The Residual Dataset

Define Common Effects Shared By the Data

Gene	Sample 1	Sample 2	Sample 3	row_mean
Gene 1	value	value	value	row_mean 1
Gene 3	value	value	value	row_mean 2
Gene 4	value	value	value	row_mean 3
col_mean	col_mean 1	col_mean 2	col_mean 3	overall mean

Gene	Sample 3	row_mean
Gene 1	value	row_mean 1
Gene 2	value	row_mean 2
col_mean	col_mean 1	overall mean

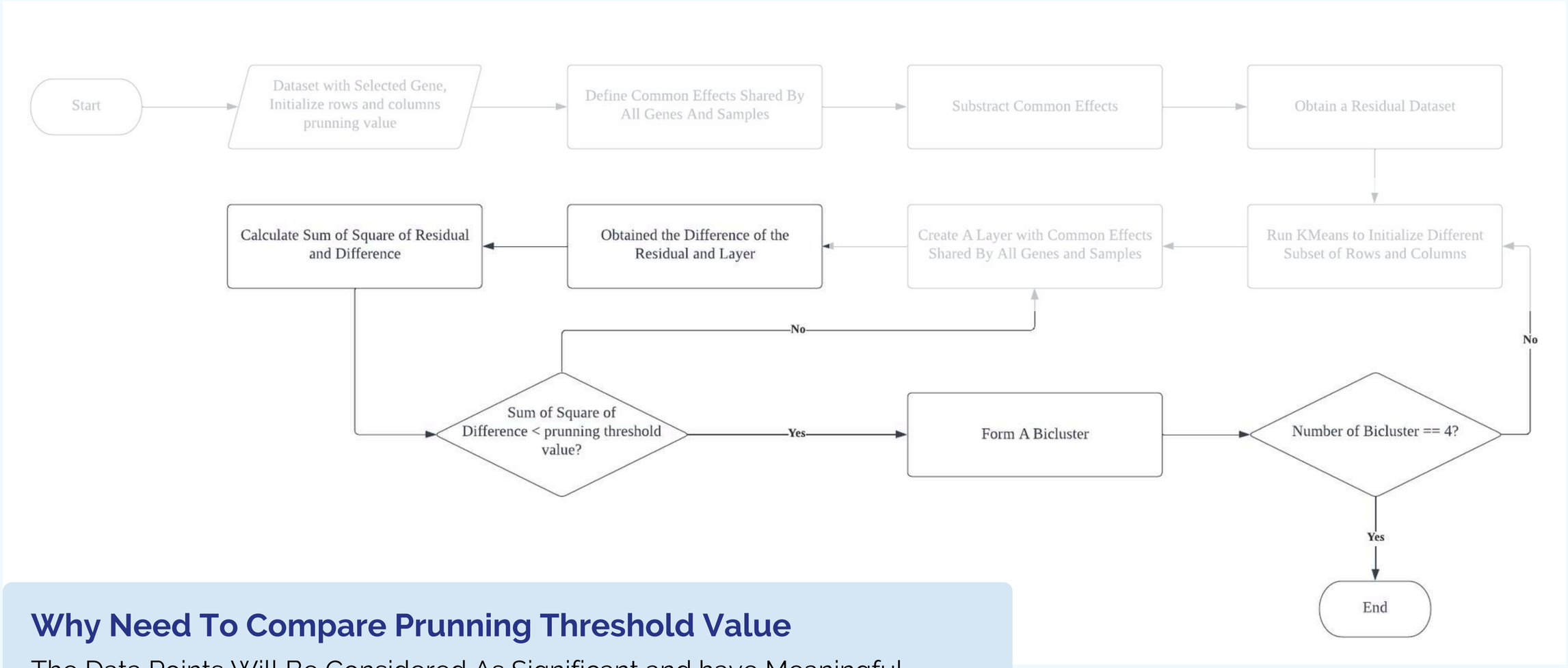
Create A Layer with Common Effects

Gene	Sample 1	Sample 2	Sample 3
Gene 1	overall mean + row_mean 1 + col_mean 1	…	…
Gene 3	overall mean + row_mean 2 + col_mean 1	…	…
Gene 4	…	…	…

Gene	Sample 3
Gene 1	overall mean + row_mean 1 + col_mean 1
Gene 2	overall mean + row_mean 2 + col_mean 1

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Plaid Model Biclustering Method



Why Need To Compare Prunning Threshold Value

The Data Points Will Be Considered As Significant and have Meaningful Pattern as the Data Points Is Close to the Expected Behaviour

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Bicluster 1

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

Bicluster 2

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

ions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Bicluster 3

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

Bicluster 4

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

Achieved Objective 2

To Implement Biclustering Algorithm in Identification of Potential Biomarkers From the Input Data of Gene Expression

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

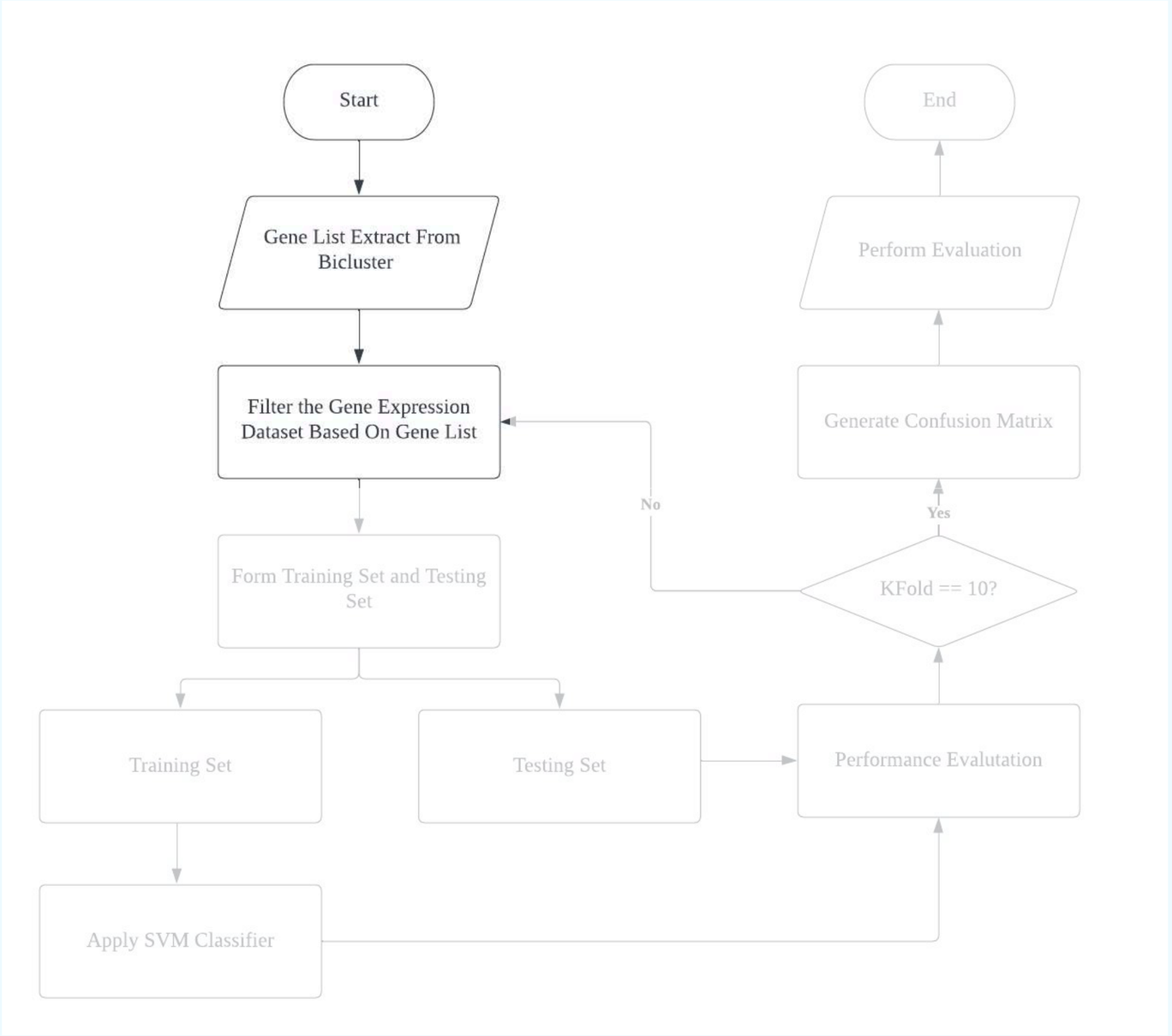
Development Process

The step to identify the potential biomarkers for EC cancer

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Performance Evaluation Process

Applying Classification Method to Generate the Confusion Matrix and Obtain the Cross Validation Value



Bicluster 1

Gene	Gene 1	Gene 2	Gene 3	Gene 4	Gene 5
Sample 1	value	value	value	value	value
Sample 2	value	value	value	value	value
Sample 3	value	value	value	value	value

Bicluster 2

Gene	Gene 3	Gene 6	Gene 7	Gene 8
Sample 1	value	value	value	value
Sample 2	value	value	value	value
Sample 3	value	value	value	value

Bicluster 3

Gene	Gene 2	Gene 9	Gene 10
Sample 1	value	value	value

Biclsuter 4

Gene	Gene 11	Gene 12
Sample 2	value	value

Gene Occur In All Bicluster

Gene 1	Gene 2	Gene 3	Gene 4	Gene 5	Gene 6
Gene 7	Gene 8	Gene 9	Gene 10	Gene 11	Gene 12

Gene Occur In Multiple Bicluster

Gene 2	Gene 3

Gene Expression Dataset that Involved Gene Occur in All Bicluster

Gene	Gene 1	Gene 2	…	Gene 12	Target
Sample 1					
Sample 2					
Sample 3					
…					

Gene Expression Dataset that Involved Gene Occur In Multiple Bicluster

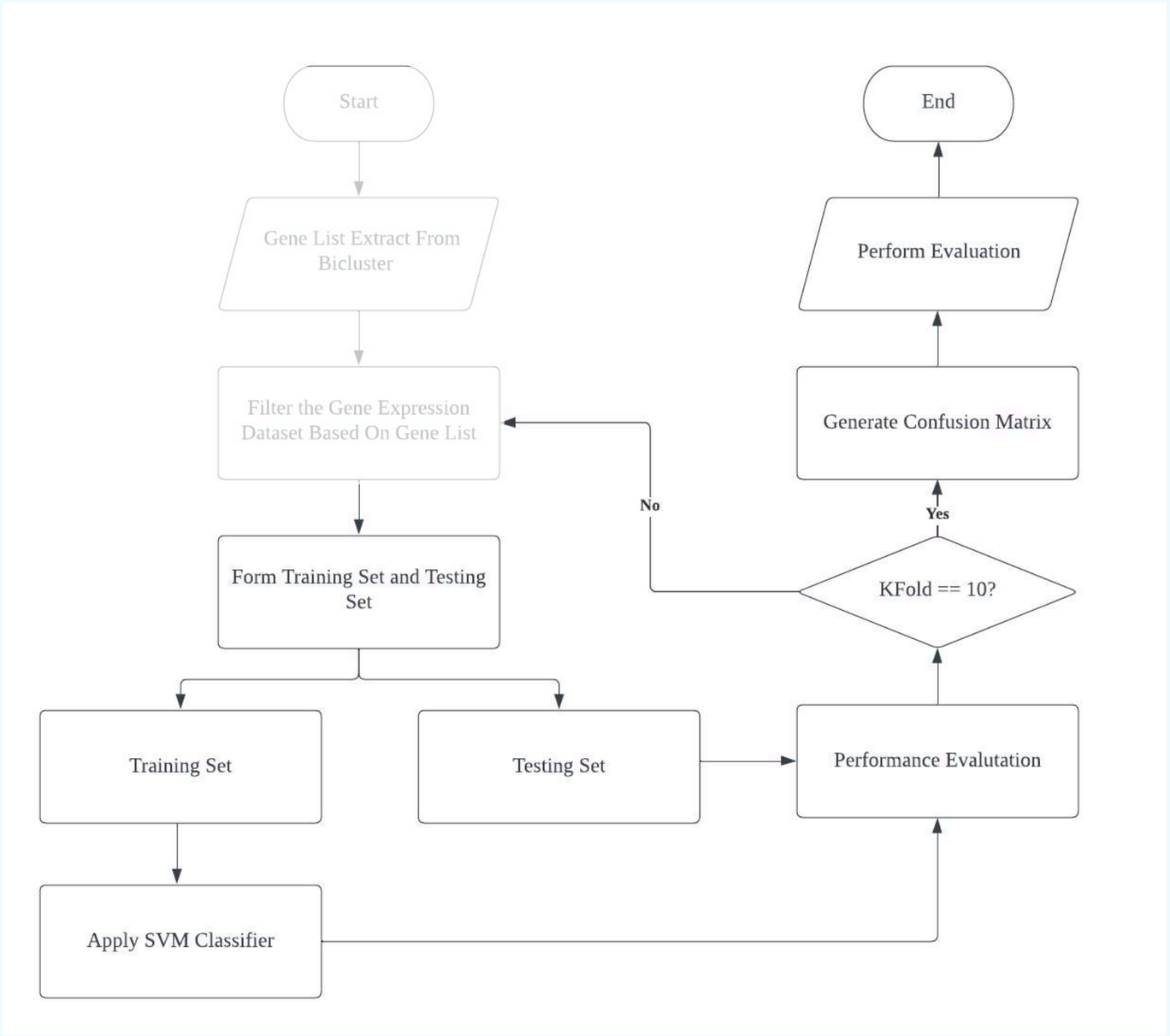
Gene	Gene 2	Gene 3	Target
Sample 1			
Sample 2			
Sample 3			
…			

Original Dataset

Gene	Gene 1	Gene 2	…	Gene 2735	Target
Sample 1					
Sample 2					
Sample 3					
…					

Purpose of Filter Out Gene Expression Dataset

To Compare the Performance Measurement from Different Subset of Data and Obtain the Potential Biomarkers that Achieved Better Result


Start
Gene List Extract From Bicluster
Filter the Gene Expression Dataset Based On Gene List
Form Training Set and Testing Set
Training Set
Testing Set
Apply SVM Classifier
Performance Evalutation
KFold == 10?
No
Yes
Generate Confusion Matrix
Perform Evaluation
End

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Performance Evaluation Result

	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

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Performance Evaluation Result

Achieved Objective 3

To Evaluate the Selected Potential Biomarkers Using SVM by Creating the Features to the Input Data of Gene Expression and Splitting them into Train and Test Data

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

Performance Evaluation Result

Achieved Objective 3

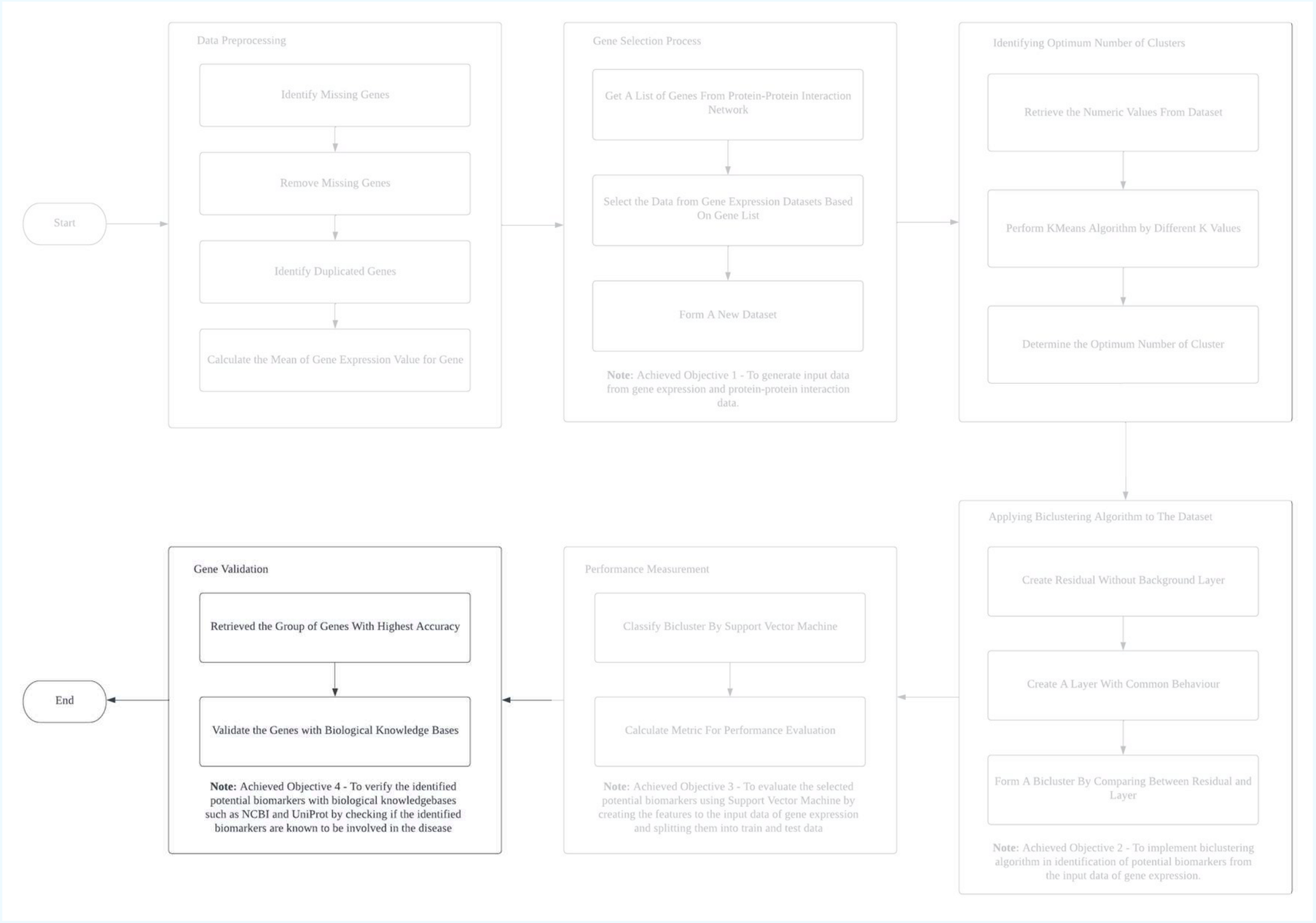
To Evaluate the Selected Potential Biomarkers Using SVM by Creating the Features to the Input Data of Gene Expression and Splitting them into Train and Test Data

Metrics	Gene Expression Dataset That Involved Genes		
	In All Bicluster	Occur In More Than One Bicluster	Original Dataset
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

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Development Process

The step to identify the potential biomarkers for EC cancer



UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Gene Validation

- Gene Expression Dataset that involved the genes occurred in multiple bicluster **achieved better result** due to the dataset had smaller sample size compared to thers
- The **frequently involvement** in multiple biclusters indicated the **consistency of genes aligned with the biological patterns** found in the data.
- **EPHB4**, **LAMB3** and **HOXD11** identified as potential biomarkers for EC Cancer

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

EPHB4

- The study on exploring the roles of cation dependent mannose 6-phosphate receptor (M6PR) and ephrin B type receptor 4 (EPHB4) in serine exosomes in promoting tumor angiogenesis and invasion of EC celss had been carried out
- Exosomes generated from EC cells that **overexpressed serine showed higher amount of EPHB4** indicating a potential role in the development of cancer
- Exosome EPHB4 increase EC cell capacity for invasion indicating a **potential function in tumor malignancy and metastasis**

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

LAMB3

- The study on assessing the expression LAMB3 in EC stem cell and adherent cell had been done.
- **Involvement of LAMB3** in the development of EC stem cells and the advancement of tumors highlight its significance as a **potential cause of the cancer**.
- Helps **produce Laminin 332**, an important extracellular matrix protein for the EC.
- **Downregulation of LAMb3** increase sphere formation, enhancing the EC traits such as **self-renewal and tumorigenicity**

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

HOXD11

- HOX gene family important for embryonic developement
- **Dysregulation** (uncontrolled growth of cell) of HOX gene disrupt the normal developmental functions causing **cancer cell to behave abnormally**
- Dysregulated of HOX gene **affect cell proliferation** (cell growth), **metastasis** (spread of cancer cell) and **treatment resistance of cancer cells**
- **Overexpression of HOXD11** can lead to **uncontrolled growth of cells**, enable the tumor to spread quickly and escape the regulatory system

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

CONCLUSION AND RECOMMENDATION

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Research Outcomes

01

Generate New Input from Gene Expression and PPI Network

02

Implement Biclustering Algorithm to form Biclusters

03

Evaluate the Potential Biomarkers by SVM Classifier

04

Verify Potential Biomarkers: EPHB4, LAMB3 and HOXD11

Innovating Solutions

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

Future Works

01

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

02

Integration of machine learning techniques to enhance the performance and scalability of biclustering algorithm in handling high dimensional dataset.

UTM
UNIVERSITI TEKNOLOGI MALAYSIA

THANK YOU

Innovating Solutions