International Journal of Innovative Computing 15(01) pp-pp

INTERNATIONAL JOURNAL OF INNOVATIVE COMPUTING
ISSN 2180-4370
Journal Homepage : https://ijic.utm.my/

Identification of Potential Biomarkers for Esophageal Cancer from Gene Expression and Interactions Using Biclustering Algorithms

Gui Yu Xuan
Faculty of Computing
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
guixuan@graduate.utm.my

Dr Chan Weng Howe
Faculty of Computing
Universiti Teknologi Malaysia
Johor Bahru, Malaysia
cwenghowe@utm.my

***Abstract*—The lack of biological relevance data in biclustering analysis leads to low precision in identifying relevant gene clusters and decreasing the accuracy of biomarker detection. The purpose of this study is to propose a biclustering method to identify the potential biomarkers of esophageal cancer (EC) from gene expression dataset and protein-protein interactions. In this research, the gene expression dataset and protein-protein interaction (PPI) data will be undergoing the gene selection process and use for biclustering method. A plaid biclustering model will be used to extract the data into several biclusters. The genes in each bicluster will be observed and filtered the data in the gene expression dataset. Then, the datasets formed were applied by Support Vector Machine (SVM) to evaluate the performance. The dataset with higher accuracy will then be validated with biological knowledge bases. The potential biomarkers found in the experiment are EPHB4, LAMB3 and HOXD11.**

Keywords — Plaid Biclustering Model, Esophegeal Cancer (EC), Potential Biomarkers, Support Vector Machine (SVM)

I. INTRODUCTION

Esophageal cancer (EC) is the world's eighth most frequent cancer [1]. Due to the 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 [2]. Gene expression data give information about the levels of gene activity but do not fully capture the complexity of biological systems [3]. Hence, 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) data. Applying conserved pathways and protein complexes, alignment and mapping of PPI networks offers a chance to learn more about the evolutionary links across species [4]. As a result, the integration of information on PPI and gene expression data enables the discovery of possible biomarkers and advances the understanding of disease.

According to the 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 [5]. Hence, by identifying biomarkers for EC have the potential to lower the morbidity and death. Biclustering is a strong data mining approach that enables grouping of rows and columns concurrently in a matrix form dataset [6]. Biclustering methods are useful for analysing gene expression and PPI data because they identify sections of genes with comparable expression pattern across sample subsets or situations [6]. Therefore, the biclustering method is a useful tool that can be used to analyze EC through the gene expression data and PPI data to detect the gene clusters that exhibit differential expression when compared to normal tissue in esophageal tumors. Besides that, biclustering can decrease the high dimensional character of gene expression datasets by focusing on their co-expressed genes, which can increase classification accuracy by decreasing the noise and highlighting pertinent features.

The lack of biological relevance data in biclustering analysis is leading to low precision in identifying relevant gene clusters and decreases the accuracy of biomarkers detection. Hence, this study is aimed to propose a biclustering method to identify the potential biomarkers of esophageal cancer from gene expression data and PPI. PPI and gene expression data are biological relevance data because they provide the interaction between genes and show the pattern of genes [7][8]. The objective of research includes to generate the input data from gene expression dan PPI data, to implement biclustering algorithm in

identification of potential biomarkers from the input data, to evaluate the selected potential biomarkers using Support Vector Machine (SVM) and to verify the identified potential biomarkers with biological knowledgebases such as NCBI. This research is aimed to contribute a biclustering method which able to identify the potential biomarkers of disease effectively. Thus, helping to the development of effective diagnostic strategies for disease.

II. Literature Review

A. Gene Expression

Data on gene expression is a measurement of the degree of gene activity in a particular cell, tissue, or organism [9]. Thus, it provide the information for medical diagnosis as the genes in the datasets are functional molecules that are involved in specific cellular process[9]. In summary, it is possible to obtain insight into the important underlying biological mechanisms and pathways by identifying the differential expression patterns of genes linked to a particular disease or condition.

B. Protein-Protein Interaction (PPI)

Essential biological procedures in cells that directly affect our healt, such as DNA replication, transcription, translation and transmembrane signal transmission, depend on proteins that have specialised functions [10]. Protein complexes, which frequently governed by PPI regulate the biological processes outlined above [10]. PPIs are essential signalling pathways in the development of various disease states, making them ideal targets for therapeutic discovery [11]. 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 [11]. As a result, PPI networks can be studied to find the relevant proteins or nodes that function as possible biomarkers and have a significant impact on cancer pathways.

C. Unsupervised Clustering Machine Learning in Biomarker Detection

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 [12]. There are total of nine methods in biclustering algorithms. The advantages and disadvantages of each method have been visualized in the table below.

TABLE I. Summary of Biclustering Algorithm

Biclustering Algorithm	Advantages	Disadvantages
Correlated Pattern Biclustering (CPB)	• Work well in synthetic dataset • Perform well in large numbers of biclusters	• Sensitive to noise • Low ability to detect higher differential expression
QUBIC	• Better Execution Time	• Low accurate and reliable result
Bayesian Biclustering (BBC)	• Well-handled missing values	• Sensitive to noise level and size
Binary Inclusion Matrix (BiMax)	• Effective for simple structure	• Sensitive to size • Limited to discrete values datasets
Plaid	• Advanced in capturing overlapped bicluster • Low coherent variance	• Sensitive to parameters used
Iterative Signature Algorithm (ISA)	• Able to find hidden homogenous group	• Sensitive to errors and outliers • Favor strong signals
Spectral	• Able to identify unique molecular subtypes • Higher enrichment analysis	• Sensitive to noise level and sample size
Order Preserving Submatrix (OPSM)	• Extract overlapped bicluster accurately • Provide stable output	• Do not filter output • Unable to analyse gene expression datasets
Cheng & Church (CC)	• Able to identify large number of bicluster	• Performance limited to higher noise level • Vulnerable to local optima • Long execution time

D. Classification Methods for Gene Expression Data

Among the classification methods used to categorize cancers are Support Vector Machine (SVM), K-Nearest Neighbours (kNN), neural networks and decision tree. Table below indicated the review on the classification methods for identifying potential biomarkers from gene expression data.

TABLE II. Summary of Classification Method

Classification Methods	Advantages	Disadvantages
SVM	• Flexible • Handle High Dimensional Datasets	• Can be expensive and complexity • Require a lot of processing power
kNN	• Simple and adaptable to noisy data • Capable of handling situations missing attributes values	• Performance based on parameter • Computationally expensive • Treats all attribute equally

Classification Methods	Advantages	Disadvantages
Neural Network	• Can capture complex relationships • Flexible	• Difficulty visualizing the decision-making process • Time consuming
Decision Tree	• Simple to visualize and analyze • Require 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

III. Research Methodology

In this chapter, a step-by-step procedure had been laid out for identifying possible biomarkers for EC, starting with data preprocessing 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.

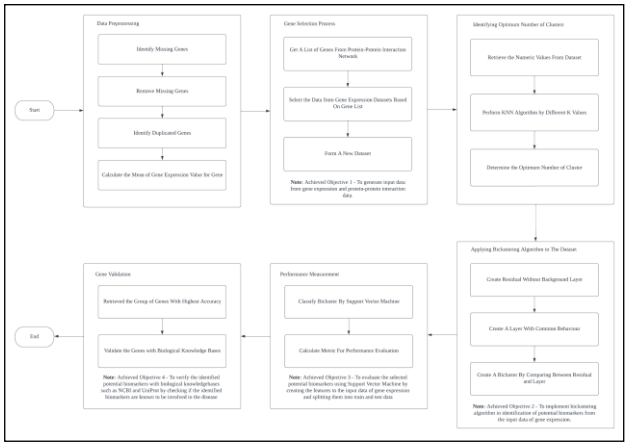

Fig. 1. Development Process

A. Data Preprocessing

There are two datasets that will be used for further analysis. Both dataset were obtained from Gene Expression Omnibus (GEO) and STRING database.The dataset obtained from GEO named as GSE20347 which contains 22278 rows of gene and 34 columns of samples where 17 of them are tumours while 17 of them are normal. There are 3506 human genes showing the connection between each other in STRING database.

In data preprocessing phase, the missing genes in the gene expression dataset were removed and eliminated to improve the computational efficiency and the accuracy result. Besides that, the gene which occurred more than one will then be calculated to obtain the average value. The dimension of the gene expression dataset after removing the missing genes and obtaining the average value of duplicated genes became 13514 genes with 34 samples.

B. Gene Selection Process

The human genes in PPI data are extracted to select the genes in the gene expression dataset. In details, the genes in the PPI data were retrieved and act as the secondary genes data. Meanwhile, the genes in gene expression dataset act as primary genes data. Then, a list of genes in the PPI data will be used to filter the data in gene expression dataset by only select those genes presented in PPI data. After gene selection process, the gene expression dataset consists only 2735 genes with 34 samples.

C. Identify the Optimum Number of Clusters

When 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 [13]. Elbow method was used to identify the optimum number of clusters due to a clear "elbow" diagram was showed from the gene expression dataset. The concept of elbow method is finding the elbow point between sum of square error and the number of clusters.

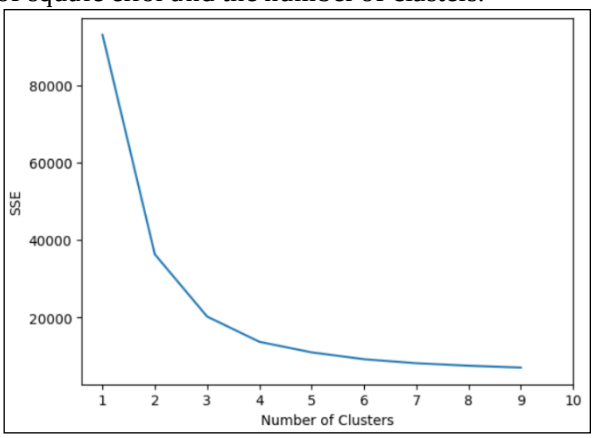

Fig. 2. Optimum Number of Cluster By Elbow Method

D. Applying Biclustering Algorithm

Figure below showed the general flows of the Plaid Biclustering Model. The plaid model can be thought as a method of breaking down the original data matrix into a collection of biclusters, each of which represent a distinct pattern in the data, and then using these patterns to reconstruct the matrix. In conclusion, the rebuilt matrix can be used to visualize the relationships between various patterns and to identify the genes or traits that each pattern most closely resembles.

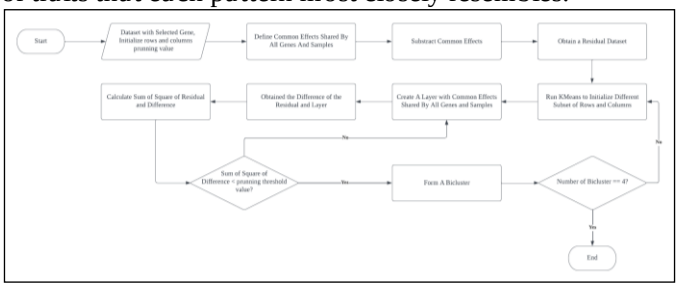

Fig. 3. Basic Architecture of Plaid Biclustering Model

1) Create Backgound Layer from Gene Expression Dataset for Pattern Capture

There is a background layer in the plaid biclustering model. Background layers indicate the common effects shared by all genes and samples. By creating a new layer, particular effects

can be separated from the background layer to show biclusters that are specific to a condition or treatment. In this step, the mean, row effects and column effects of the gene expression dataset will be calculated. This method capture both the overall behaviour of the row and column.

2) Subtracting Background Layer

By subtracting the background layer from the gene expression dataset, the algorithm effectively removed the common impact represented by the background layer. This procedure updated the gene expression dataset to concentrate on any remaining precise changes that the background layer is unable to account for. In a nut shell, after gene expression dataset subtracted the background layer, the residual is formed and the residual is important to select the data points to form the bicluster.

3) Formed A Collection of Bicluster

The process required finding coherent groups of genes and samples, capturing their shared behaviour in a common layer, verifying that the behaviour is meaningful, and finally classifying these groups as biclusters if certain criteria are met.

The first step was started by run K-Means to initialize the rows and columns. K-Means is a method to define the dissimilarity between the data points. As a result, expression patterns in genes and samples are comparable. K-Means algorithm effectively breaking up the gene expression dataset into smaller parts with similar features. At this stage, the rows and columns for new layer can be initialized.

After rows and columns are initialized, a new layer is formed. This new layer is constructed by averaging the residual and combining with the row and column effects. Essentially, this layer provides the overall behaviour observed in the subset of gene expression dataset.

In order to ensure that the common behaviour observe are meaningful rather just random occurrences, the variants in the residual and the layer had been compared. The differences of residual and layer was calculated. If teh pruning threshold value is much higher than the sum of square of the differences, it means that the layer does capture enough diversity in the data and is significant for further analysis.

IV. Performance Measurements

After obtaining the bicluster data, a SVM classifier will be used to discover the potential EC cancer biomarkers. However, the biclustering result indicated that the retrieved sample is cancerous. Since the target class only includes cancer cases, hence the data unable to undergo classification directly. However, the goal of biclustering algorithms is to discover the key features by showing patterns in gene and sample data. Thus, it is possible to assume that the genes found inside the bicluster are important indicators that may be predictive of EC. It indicated that these genes exhibit patterns that imply their involvement in disease. Hence, the genes that found inside the bicluster will be extracted to filter the gene expression dataset. Furthermore, there are some genes appear in more than one bicluster. To improve the classification results, these genes were used to extract the data from gene expression dataset for classification process. In summary, there are three gene expression dataset will be used to develop SVM classifiers. The three gene expression datasets are gene expression dataset that involved genes in all biclusters, gene expression dataset that involved genes occurred in more than one bicluster and original gene expression dataset.

Gene expression dataset that involved genes in all biclusters is constructed with the data in gene expression dataset by combining all the genes that appearred in the biclusters. Gene expression dataset that involved genes that occurred in more than one bicluster is constructed with the data in the gene expression dataset by finding the same gene that occurred in different biclusters. Original gene expression dataset is the gene expression dataset after the data preprocessing and gene selection process.

A. Apply SVM Classifier to the Gene Expression Dataset

The dataset is split into features and target variables in order to apply the SVM classifier. Subsequently, the dataset is divided into training and testing set with a 60 percent training data and 40 percent testing data ratio. The performance of a linear SVM classifier is evaluated using a ten-fold cross validation technique. A confusion matrix is also created to evaluate the classifier's performance. Furthermore, different performance metrics such as accuracy, precision, recall, specificity and F1 score are used to evaluate the classifier's effectiveness.

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 [14]. True Positive (TP) is the number of correctly predicted positive instances [15]. False positive (FP) is the number of incorrectly predicted positive instances [15]. True negative (TN) is the number of correctly predicted negative cases while false negative (FN) is the number of incorrectly predicted negative instances [15]. Accuracy is the percentage of accurate predictions the model makes is measured [15]. Precision is the ratio of accurate positive predictions to all positive predictions made by the model [15]. Recall is the ration of accurate positive predictions to all positive cases [15]. F1 score is a measurement for evaluating the overall performance by providing the balance between of precision and recall [16].

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (1)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F1 Score} = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \quad (1)$$

The tables below indicate that gene expression dataset involved genes that occurred in more than one bicluster achieved the highest accuracy and performance than others. 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 biclusters and original gene expression dataset achieved approximately 67 percent in fold six while gene expression dataset that involved genes that occurred in more than one bicluster achieved 75 percent in fold three. Besides that, SVM classifier can correctly predict the target label for 97.06 percent

of the samples in each dataset. For the gene expression dataset that involved genes in all biclusters and original gene expression dataset able to make the positive prediction and identified all actual negative target correctly as precision and specificity values as 1 while these two datasets only able to identify 94.12 percent of the actual positive target. Meanwhile, the gene expression dataset that involved genes in multiple bicluster had the precision and specificity value as 0.9444 and 0.9412 respectively and able to identify all actual positive target with 100 percent recall value.

TABLE III. PERFORMANCE EVALUATION OF GENE EXPRESSION DATASET BASED ON TEN-FOLD CROSS VALIDATION

	Gene Expression Dataset that Involved Genes		
	In All Biclusters	**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 IV. PERFORMANCE MEASUREMENT OF GENE EXPRESSION DATASET BASED ON CONFUSION MATRIX

	Gene Expression Dataset that Involved Genes		
	In All Biclusters	**Occur In More Than One Bicluster**	**Original Dataset**
Cross Validation			
Accuracy	0.9667	0.975	0.9667
Confusion Matrix			
Accuracy	0.9706	0.9706	0.9706
Precision	1	0.9444	1
Recall	0.9412	1	0.9412
Specificity	1	09412	1
F1 Score	0.9697	0.9714	0.9697

B. Applying Different Random State to Measure the Accuracy

The ten-fold cross validation results for the gene expression dataset indicated that the accuracy of each fold varied significantly. This variation implied that the model's performance is sensitive to specific data splits. To overcome these issues, total of ten different random states had been applied to the SVM classifier to ensure that the model's performance is not excessively dependent on a certain random split of the data. By evaluating a model's accuracy across numerous random states, a more reliable measurement of its performance can be obtained.

Table below demonstrated the accuracy of gene expression dataset with different random state. The gene expression dataset that involved genes in all biclusters and the original gene expression dataset achieved 96.43 percent accuracy respectively. Meanwhile, the gene expression dataset that involved genes which occurred in more than one bicluster achieved 95 percent accuracy. A conclusion of gene expression involved genes that occurred in more than one bicluster achieved the higher accuracy and better result than others can be made. This is because the sample size of gene expression dataset that involved genes that occurred in more than one bicluster is smaller than the gene expression dataset that involved genes in all biclusters and the original gene expression dataset.

TABLE V. ACCURACY OF GENE EXPRESSION DATASET WITH DIFFERENT RANDOM STATE

Random State	Gene Expression Dataset that Involved Genes		
	In All Biclusters	**Occur In More Than One Bicluster**	**Original Dataset**
5	1	1	1
10	0.9286	1	0.9286
15	0.9286	0.9286	0.9286
20	1	0.9286	1
25	1	0.9286	1
30	0.9286	0.9286	0.9286
35	1	1	1
40	0.9286	0.9286	0.9286
45	1	0.9286	1
50	0.9286	0.9286	0.9286
Average	0.9643	0.95	0.9643

TABLE VI. SUMMARY OF ACCURACY BY THE DIMENSION OF DATASET

Gene Expression Dataset of Genes	**Samples**	**Genes**	**Accuracy**
In All Bicluster	34	285	96.43%
Occur In More Than One Bicluster	34	3	95%
Original Dataset	34	2735	96.43%

C. Verify the Selected Potential Biomarkers

Even though from the above result, the gene expression dataset that involved genes occurred in multiple bicluster achieved the better results on classification but the performance across three datasets was almost the same. To further explore the relationship between three datasets, a t-test with significance

level of 0.05 was conducted on different random state accuracy values of three datasets. Figure below indicated the result of t-test. Since the p-values are higher than significance level, hence there is no significant difference between three datasets. The lack of significant differences between three datasets indicates the genes included in each dataset are likely to contribute to the EC. However, there are three genes, EPHB4, LAMB3 and HOXD11 occur in multiple biclusters which means that have the higher chances as the potential biomarkers for EC. This is because frequent involvement in multiple biclusters indicates that these genes consistently align with the biological patterns found in the data. Nonetheless, statistical result alone is not sufficient to conclusively identify biomarkers for EC. Therefore, genes in bicluster will be further validated using a biological knowledge base to ensure their relevance and importance in the EC context.

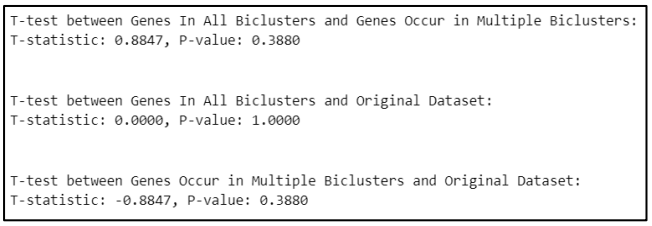

Fig. 4. T-Test Result

V. Gene Validation

A. EPHB4

A study on exploring the roles of cation-dependent mannose 6-phosphate receptor (M6PR) and ephrin B type receptor (EphB4) in serine (SRGN) exosomes in promoting tumour angiogenesis and invasion of EC cells had been carried out. Based on the findings, exosomes generated from EC cells that overexpressed SRGN showed higher amounts of EPHB4, indicating a potential role for this protein in the development of cancer [17]. Significantly, exosome EPHB4 increased EC cell's capacity for invasion, indicating a potential function in tumour malignancy and metastasis [17]. Furthermore, the significant association between EPHB4 expression and SRGN levels in EC patients' serum highlights its potential as a prognostic indicator, with high serun EPHB4 being associated with lower overall survival [17].

B. LAMB3

A study on the assessing the expression of LAMB3 in EC stem cell and adherent cells had been done. The study suggested that the involvement of LAMB3 in the development of EC stem cells and the advancement of tumours highlights its significance as a potential cause of the cancer [18]. The different expression pattern of EC stem cells and adherent cells showed that it is involved in critical processes such as spheroid formation, EC stem cell development and tumour growth [18]. LAMB3 helps to produce Laminin-332, an important exracellular matrix protein for the EC stem cell microenvironment [18]. Downregulation of LAMB3 in EC has been linked to sphere formation, implying a role is enhancing EC stem cell traits such as self renewal and tumorigenicity [18].

C. HOXD11

HOX gene family is important for embryonic developemnt and its dysregulation is association with several malignancies, including EC [19]. When HOX genes are dysregulated, their normal developmental functions are disrupted, causing cancer cells to behave abnormally [19]. Dysregulated HOX genes, such as HOXD11, may affect cell proliferation, metastasis, and treatment resistance of cancer cells, thereby affecting tumor progression and patient prognosis [19]. Dysregulation of HOX genes can have a significant impact on cancer biology [19]. 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 preven excessive cell division [19]. Dysregulated HOX genes also help cancer cells spread to distant regions of the body and improve their capacity for metastasis [19]. As a result, misregulation of HOX genes enhances the complexity of cancer development and creates major difficulties for the treatment [19].

VI. Conclusion

In conclusion, the involvement of PPI data with gene expression dataset enhances the classification performance. The result suggest that the importance of combining multiple data sources to obtain more comprehensive biological insights, leading to more accurate and robust biomarker identification. Besides that, the implementation of plaid biclustering model able to recognize the similar expression patterns leads to the grouping of genes and samples into biclusters. The plaid biclustering model searches for subset inside the gene expression dataset where the rows and columns have consistent expression patters showing the data's biological complexity and variabilty. By implement the biclustering algorithm, the chance to identify the important genes that contribute to EC cancer is high. However, the presence of noisy data, outliers and different scaling across the dataset may misrepresent the model's ability to learn the meaningful patterns. 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. The future works on this research are develop a method of determining the optimal pruning threshold value to be used in plaid model biclustering algorithm and integration of machine learning techniques to enhance the performance and scalability of biclustering algorithms in handling high dimensional dataset.

Acknowledgment

In preparing this thesis, I was in contact with many people, lecturers and friend. They have contributed towards my understanding and thoughts. 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 literature.

My fellow undergraduate student should also be recognized for their support. My sincere appreciation also extends to all my course mates and others who have aided me on 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 members.

REFERENCES

[1] 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).

[2] 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.

[3] 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.

[4] 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.

[5] 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).

[6] 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.

[7] 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.

[8] 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).

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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)

[14] 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.

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

[16] 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)

[17] 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.

[18] 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.

[19] 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.

[20]