IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Ingénierie des Systèmes d'Information
Vol. 27, No. 3, June, 2022, pp. 487-493
Journal homepage: http://iieta.org/journals/isi

Differential Evolution Model for Identification of Most Influenced Gene in Brest Cancer Data

Hymavathi Thottathyl*, Kanadam Karteeka Pavan

Department of Computer Applications, R.V.R.& J.C. College of Engineering, Chowdavaram, Guntur 522019, A.P., India

Corresponding Author Email: hyma@rvrjc.ac.in

https://doi.org/10.18280/isi.270316

Received: 17 February 2022
Accepted: 15 April 2022

Keywords:
microarray data, clustering, unsupervised learning, unlabeled data, gene expression

ABSTRACT

Microarray technology generates a large amount of data. Clustering is a popular technique for locating genes that are expressed in close proximity. It entails examining a fresh dataset to determine whether similar traits can be used to identify any hidden groupings. There are large-dimensional datasets accessible, such as those produced from gene expression investigations, RNA microarray studies, or RNA sequencing studies. As a consequence, cluster analysis and producing well-separated clusters become more challenging. Good cluster separation is desirable since it suggests that items are not being placed in the erroneous clusters. In this study, it was recommended that a Differential Evolution-based (DE) Model be used to interpret the analysis of Brest cancer gene expression. To begin, cluster the gene expression data to find the genes most likely to be impacted by the illness. The appropriate number of clusters must be found in order to locate the gene with the highest effect in the gene collection. We used a DE model on the Brest cancer datasets in this work. We identified the best number of clusters by using the most impacted gene in this dataset as a benchmark. We then experimented with different cluster sizes. We used the DE method to three distinct breast cancer datasets and compared it to the current K-Means and K-medoids models. The results of the experiments show that the proposed DE model outperforms existing models significantly.

1. INTRODUCTION

When Hippocrates first discovered cancer in the fifth century BC, the word "cancer" was translated into Latin as "crab." Eventually, as cells divide and spread uncontrolled throughout the body, they smother the organism and cause it to die. At least many thousand years have elapsed since the sickness initially manifested itself, and its generality has been regularly enlarged during that period. Cancer seems to attack elderly individuals at a higher rate than it does small people, and cancer cases are increasingly being discovered in persons throughout 65, according to research. According to the WHO, increasing access to vaccinations and antibiotics is the primary cause of an increase in the average lifespan, which have decreased the mortality rate from infectious diseases, has increased the number of people living long enough to develop cancer. Worldwide, the reason for women to die is A form of cancer that affects the breasts, and its effect is the same in men and women. The cause of mortality in women is breast cancer globally; every year, 1.4 million cases are diagnosed in middle-income and low-income countries, accounting for more than half of the world's four lakhs new breast cancer fatalities. However, even though cancer-specific mortality has been decreased due to therapeutic procedures such as surgery, chemotherapy, and radiation treatment, there are still more therapeutic failures that occur in the recurrence of cancer, metastasis, and mortality in cancer patients. In terms of breast cancer, sex and age are the primary risk factors. the incidence of breast cancer in women is about 100 times greater than in men, and approximately 80 percent of the cases are observed in women at the age of 55. It is estimated that breast cancer is caused due to genes of the family around 10%, which is generally caused by a hereditary mutation in one of two genes, the BRCA1 or BRCA2. Breast cancer is more likely in women who have undergone significant ionizing radiation in the chest region. The condition is influenced by a woman's reproductive and hormonal history and genetic makeup. Those women who begin menstruation at a young age or have menopause at a late age and those who have never given birth or according to research, women who have had children beyond the age of 30 have a slightly increased risk of having breast cancer. According to the findings of many research conducted so far, women who take oral contraceptives have a modest increase in their chance of getting breast cancer. Additional risk factors include using alcoholic drinks, being overweight, and engaging in little physical exercise. According to some findings, the environment with pesticides that act as estrogens may increase the chances of breast cancer. On the other hand, the present research does not consistently support this notion. Most of the time, a lump identified during a normal breast self-exam is the first indication of breast cancer. In addition to these symptoms, additional possible markers are unexplained breast expansion and thickness, skin irritation, discomfort, nipple soreness, and the discharge that is not breast milk. Using mammography, it is possible to detect small breast tumours that the patient or her doctor might otherwise go undetected.

It is now recommended that all females are older than 40 [1] and the beginning period of 40 get frequent mammograms to screen for breast cancer at the earliest opportunity. A biopsy is conducted to determine whether cancer has been established in the tissue whenever a suspicious lump is discovered by physical examination or mammography. Several variables have been linked to the development of breast cancer, but the precise mechanism by which the disease manifests itself is still unclear. Before trying to decipher this process, it is essential to have a thorough grasp of the genes involved and their regulated mechanisms. It is now feasible to measure gene expression across the entire genome at a single time point using DNA microarray technology, which was previously impossible. We may learn more about gene interactions and how this affects the pathophysiology of diseases and the course of such illnesses by looking at gene expression patterns. In 2002, M. Fey and colleagues released a study. DNA microarrays have been utilized to explore the underlying biology of various cancer forms, most notably breast cancer. It is possible to use this sort of study to examine all aspects of the illness, from its onset and course through its invasion and drug resistance, to identify the underlying molecular mechanisms involved in each of these processes. It can also describe the disease's beginning, progression, invasion, and medication resistance. Various studies have been conducted on breast cancer cell lines utilizing DNA microarrays to generate molecular profiles, classifications, and prognostic markers for breast cancer. As a result of this research, Breast cancer genes, such as those involved in tumour suppression and metastasis inhibition, will likely be found, and we'll be able to learn more about how they affect the activity of many other genes and pathways across the body. When these genes are controlled, we may be able to cure cancer or at least limit its spread. Genetic ontology study was carried out utilising Gene spring's array analysis in comparison to Gene spring's array analysis in the control sample. As a consequence of these evaluations, two genes have been selected for future study in terms of both their regulation and their relationship to cancer. Quantitative real-time PCR confirmed the presence of these two genes in the MDA-MB-231 breast cancer cell line. using the reference gene-actin as a control. (Which is a housekeeping gene) as a housekeeping gene. Genes involved in tumor invasiveness may be discovered due to this research. Figure 1 explains proposed model architecture.

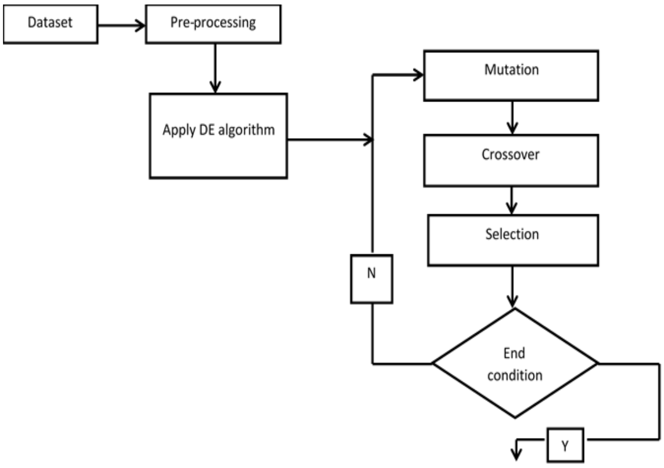

Figure 1. System architecture

We suggested a Differential evolution-based approach for analyzing Brest cancer gene expression data in this research. First, take the gene expression data and determine which gene is the most influenced by it. Identify the genes that are most impacted by the Brest cancer data. These data were analyzed using several datasets, and they performed comparisons. Here are the sections of the paper: The literature review is summarised in sections 2–5, the DE model for gene expression data interpretation is Section 3 explains the experiment, while Section 4 shows the results, and Section 5 concludes the paper.

2. LITERATURE SURVEY

“Classification of Microarray Gene Expression Data Using an Infiltration Tactics Optimization (ITO) Algorithm." It was written by Zahoor and Zafar [2]. The ITO algorithm has a combination of parameter-free and parameter-based classifiers. Higher accuracy doesn't mean that higher reliability.

"Microarray cancer feature selection: Review, challenges, and research directions." It was written by Hambali et al. [3]. The selection of features and dataset classification is used. The various components of Feature selection are explained in a detailed manner. Feature selection is critical and sensitive.

"Model-Based Modified K-Means Clustering for Microarray Data." It was written by Suresh et al. [4]. The Novel Clustering is used for the Clustering of microarray data. The EVV is the best model with three components, and it helps developers think and fix the greater number of clusters. The limitation is, we need to tell the number of clusters.

"Dealing with high-dimensional class-imbalanced datasets: embedded feature selection for SVM classification." It was written by Maldonado and Lopez [5]. It uses the algorithm with have two SVM formulations. It can deal with an imbalance of class and more dimensionality issues. The datasets used are binary class data.

"An Adaptive Harmony Search Approach for Gene Selection and Classification of High Dimensional Medical Data." It was written by Dash [6]. It uses hybridized harmony search. It is used to draw a combination of harmony searches.

"Gene Selection for Microarray Cancer Classification Using a New Evolutionary Method Employing Artificial Intelligence Concepts." Dashtban and Balafar [7] wrote it. It uses genetic algorithms which are based on the new evolutionary approach. For feature selection, a hybrid of GA is developed. Most of the datasets have fewer features, i.e., 10000, and there is no similarity.

"Novel Hybrid Hierarchical-K-means Clustering Method (H-K-means) for Microarray Analysis." It was written by Chen et al. [8]. In one cycle, we cluster half of the data using hierarchical Clustering and the other half using K-means. Initiation is K-means; it's a good idea to mix the two procedures. If we stop hierarchical Clustering between 40% and 60%, the outcomes are better.

"Clustering-based Spot Segmentation of cDNA Microarray Images." It was written by Uslan and Bucak [9]. We use fuzzy c-means and clustering-related methods for the segmentation step, separating the spots from the background.

"Determination of the Minimum Sample Size in Microarray Experiments to Cluster Genes Using K -means Clustering." It was written by Wu et al. [10]. It uses the K-Means clustering method. Accurate results can be achieved with minimum cost. The dataset needed is vast; we have to filter it.

"Automating Microarray Classification using General Regression Neural Networks." It was written by Soares et al. [11]. It uses AUROC and Accuracy. The results are better when microarray classifications are used than other techniques. There is no optimal parameter for PSO.

"K-means+ Method for Improving Gene Selection for

Classification of Microarray Data." It was written by Huang et al. [12]. It uses Chebyshev's Theorem and Empirical Rule are used. Clustering methods are used to reduce redundancy. The no. of clusters must be given, and this process consumes more time to try all possible numbers for clusters.

"Clustering Microarray Data using Fuzzy Clustering with Viewpoints." It was written by Katerina et al. [13]. we combine the external domain language with fuzzy Clustering, and viewpoints represent it. These results are better when compared to the first three clustering algorithms. To calculate the viewpoints, we should find the average value of each feature.

"A Principal Component Analysis Based Microarray Data Bi-clustering Method." It was written by Zhang et al. [14]. Biclusters can be identified directly with the whole experiment data matrix. We use conventional clustering methods. The combination of biclusters is complicated when it comes from different types.

"A Combined K-Means and hierarchical clustering methods for improving the clustering efficiency of microarray." It was written by Chen et al. [15]. We use an agglomerative approach. Bottom-Up Clustering is slow when compared to Top-Down clustering. Divide Hierarchical K-Means have results accurately and fastly.

"The Application of an Improved K-means Clustering Method in Microarray Gene Expressing Data." It was written by Guan et al. [16]. It uses the better method of traditional K-means clustering. It predicts the fit no. of clusters and draws the cluster. The regulation of genes is a highly complex process.

The goal of gene expression clustering is to identify and extract cohorts of genes that are behaving in a coordinated manner from the complete set of genes that are detected within a particular context, rather than to partition the complete set of genes into a set of gene clusters as previously stated. The previous studies made analysis on gene expression analysis but most of the methods are not made extensive analysis with different number of clusters to get optimal clusters.

3. PROPOSED WORK

It has been shown that clustering approaches are useful in understanding genes and their functions, gene control, cellular processes, and cell subtypes. It is possible to group genes with comparable expression patterns (co-expressed genes) with genes that perform similar biological activities. It is possible that this technique will contribute to a better understanding of the activities of numerous genes for which information has not previously been accessible. Moreover, genes that are co-expressed in the same cluster are more likely to be engaged in the same biological activities, and a significant connection between the expression patterns of those genes implies that they are regulated by one another. The identification of regulatory motifs unique to each gene cluster and the proposing of cis-regulatory elements [10] that are particular to each gene cluster may be achieved by searching for comparable DNA sequences at the promoter regions of genes within the same cluster. Gene expression data may be used to make inferences about the mechanism of the transcriptional regulatory network, based on clustering [17, 18], which are discussed further below. In the end, grouping distinct samples based on their matching expression patterns may uncover sub-cell types that are difficult to detect using classic morphology-based techniques. DE algorithm is applied on breast cancer gene expression data. Initially take gene expression dataset and applied DE model and repeatedly made clustering from two to ten clusters and identify most convergence clusters.

DE is an Evolutionary Algorithm
Consider an optimization problem
Minimize f(Y)
where, $Y = [y_{1,} y_{2,} y_{3,\ldots.,} y_D]$, D is the number of variables.

Algorithm:
Step 1:
The population matrix can be shown as:

$$y^g{}_{m,j} = [y^g{}_{m,1}, y^g{}_{m,2}, y^g{}_{m,3}, \ldots., y^g{}_{m,D}]$$

where, g denotes the Generation and m=1, 2, …, M.

Step 2: Initialisation

$$y_{m,j} = y_{m,j}{}^v + rand(\,) * (y_{m,j}{}^u - y_{m,j}{}^v)$$

j=1, …., D and m=1, ..., M.
where, $y_j{}^v$ is variable $y_j$lower bound.
$y_j{}^u$ is variable $y_j$lower bound.

Step 3: Mutation
Select the three other vectors $y_{r1m}{}^g, y_{r2m}{}^g, and\ y_{r3m}{}^{g\ from\ every\ parameter\ vector}$.

$$vd_m{}^{g+1} = y_{r1m}{}^g + F(y_{r2m}{}^g - y_{r3m}{}^g)$$

m=1, 2, 3, …, M,
$vd_m{}^{g+1}$ is referred as donor vector,
F is had values between 0 and 1.

Step 4: Recombination

$$t_{m,j}{}^{g+1} = \begin{cases} vd_{m,j}{}^{g+1}\ if\ rand(\,) \leq r_p\ or\ j = J_{rand} & j = 1,2,3, \ldots D \\ y_{m,j}{}^g\ if\ rand(\,) > r_p\ and\ j \neq J_{rand} & m = 1,2,3, \ldots M \end{cases}$$

J_{rand} is a random integer number between [1, D];
r_p refers to recombination probability.

Step 5: Selection

$$y_m{}^{g+1} = \begin{cases} t_{m,j}{}^{g+1} & if\ f(t_m{}^{g+1}) < f(y_m{}^g) \\ y_m{}^g & otherwise \end{cases}$$

m=1, 2, 3, …, M.

4. EXPERIMENTAL RESULTS & DISCUSSIONS

Datasets:

Cancers are related to anomalies in the genetic code. Gene expression measures the degree of activity of the gene in the tissue and provides data about the complicated actions carried out by the gene. By comparing the standard and sick tissue genes, researchers may get more insight into the prognosis and fate of cancer patients. ML algorithms to genetic data can provide an accurate estimate of survival time while also avoiding the need for unneeded surgery and therapeutic operations.

For 331 genes, z-scores and mutations for 175 genes are among the data provided by the collection's genetics section. With the use of mRNA expression data, it is possible to estimate the tumor's relative expression in relation gene expression patterns in a healthy population. The "reference population" consists of all of the samples that were utilised in the study. Respondents are shown how many standard deviations they are from the reference population's mean expression. (Z-score). It is possible to determine if a gene is more or less expressed in a cancer sample when compared to a control sample using this approach. A Canada-United Kingdom project has produced the METABRIC database, which comprises targeted sequencing data from 1,980 primary breast cancer samples.

These factors include genomic and transcriptomic as well as epigenetic components that are involved in the genesis and development of breast cancer." It includes transcriptome gene expression and somatic mutations from the TCGA BRCA. More than 57,000 genes are included in the BRCA gene expression collection, which consists of 1222 samples. In all, 1109 tumour samples and 113 reference samples have been gathered for the study's examination. By utilising the edgeR algorithm, we eliminated any genes whose expression was only moderately elevated across most samples, and then we normalised our results. The number of genes with low levels of expression in the majority of samples was reduced from 57,063 to 34,465 as a consequence of the deletion.

It is hoped that additional participants and more precise phenotypic labelling will lead to better models for predicting breast cancer risk. SNPs from a large cohort of people were utilised to develop classifiers that can predict if a new person is prone to breast cancer (cases and controls).

In a GWPS, there are N participants (cases and controls) who had their SNP profiles analysed. In order to develop a classifier for new individuals, it uses an individual's SNP profile.

Only 348 breast cancer cases and 348 control subjects were recruited for the research, all of whom were Caucasian and had no family history or risk factors for breast cancer when they were included. Affymetrix Human SNP 6.0 arrays were used by the researchers to analyse the data. A previous study on sporadic breast cancer found that women with a family history of the illness were more likely to get the disease themselves.

DUNN Index:

Using the Dunn index (DI) to evaluate clustering techniques, the outcome is based only on the clustered data. Each of these indices has the goal of identifying clusters that's well, with low variation between the member nodes and a significant distance between both the means of various clusters, compared to the variance within the clusters.

The greater the grouping, the higher the Dunn index. The ideal number of clusters k is the number of clusters that maximise the Dunn index. Isn't there a downside? Data complexity and cluster size go hand in hand, with rising costs.

DI for c no. of clusters is defined as:

$$Dunnindex(U)=\left\{\min_{1\le i\le c}\left\{\min_{1\le j\le c, j\ne i}\left\{\frac{\delta\left(X_i,X_j\right)}{\max\limits_{1\le k\le c}\left\{\Delta\left(Xk\right)\right\}}\right\}\right\}\right\}$$

where,

$\delta(X_i,X_j)$ is the distance between inter clusters.

$\Delta(Xk)$ is the distance between intra clusters.

Silhouette coefficient:

When we talk about silhouettes, we're talking about a means of interpreting and validating consistency inside clusters of data. With this method, you can quickly see how successfully each item has been categorized via a brief graphical depiction. The silhouette value, which is derived using the silhouette technique, measures how similar an item is to its cluster compared to other clusters.

$$S(i)=(b(i)-a(i))/(\max\{(a(i), b(i))\})$$

where,

•Dissimilarity between ith item and other objects in its cluster averages out to a (i).

•b(i) is the average degree to which an item in a cluster differs from all the others.

DB Index:

According to David Davies and Donald W. Bouldin, the DBI (Davis–Bouldin index) is an internal assessment method for the performance of clustering algorithms. Verifying how far the procedure performed is based on data quantity and characteristics. Clustering is better when the DB index value is smaller. It has its downsides, too. There is no guarantee that a high value is indicative of the best information retrieval when using this strategy.

DB index for k no. of clusters is defined as:

$$DBindex(U)=1/k\sum_{i=1}^{k}\max_{i\ne j}\left\{\frac{\Delta(Xi)+\Delta(Xj)}{\delta\left(X_i,X_j\right)}\right\}$$

where,

$\delta(X_i,X_j)$ is the interclassed distance.

$\Delta(Xk)$ is the intracultural distance.

Here Figure 2 describes the DUNN index comparison analysis with K-means, k-medoids, and proposed DE models applied on the METABRIC dataset. Number of clusters on the x-axis is shown, and the y-axis has the DUNN index. DUNN index of the proposed model shows good compared to existing modes by varying the number of clusters. The dataset contains 331 genes information. The model clusters the genes causing the breast cancer data. Here DE model is converged at eight cluster it gives DUNN index nearly 0.98.

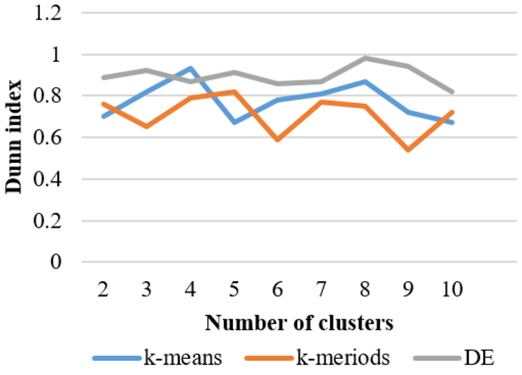

Figure 2. DUNN index on METABRIC data

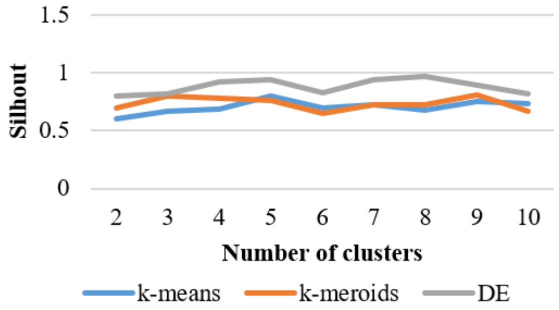

Figure 3. Silhouette index on METABRIC data

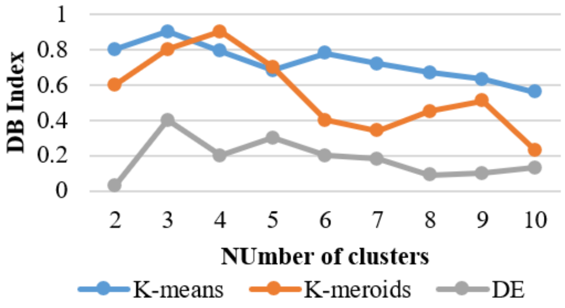

Figure 4. DB index on METABRIC data

Here Figure 3 shows the Silhouette index comparison analysis with K-means, k-medoids, and proposed DE models applied on the METABRIC dataset. Number of clusters on the x-axis is shown, and the y-axis has the Silhouette index. Silhouette index of the proposed model shows good compared to existing modes by varying the number of clusters. The dataset contains 331 genes information. The models cluster the genes that cause the breast cancer data. The model clusters the genes causing the breast cancer data. Here DE model is converged at eight cluster it gives Silhout index nearly 0.99.

Here Figure 4 shows the DB index comparison analysis with K-means, k-medoids, and proposed DE models applied on the METABRIC dataset. Number of clusters on the x-axis is shown and the y-axis has the DB index. DB index of the proposed model shows good compared to existing modes by varying the number of clusters. The dataset contains 331 genes information. The model clusters the genes causing the breast cancer data. The model clusters the genes causing the breast cancer data. Here DE model is converged at eight cluster it gives DB index nearly 0.09.

The three results Figures 2-4 are representing DUNN Index, Silhouette Index, DB Index on METABRIC data respectively. Clusters are varying from two to ten clusters. By observing the clusters at eight clusters we found better convergence in the DE model results. That is DUNN index is 0.98, Silhout index is 0.97 and DB Index is 0.09. These results indicated in METEBRIC data the most influenced gene has been found after making eight clusters in METABRIC dataset. Other models are failed to make proper convergence and results not promising like DE model.

Here Figure 5 shows the DUNN index comparison analysis with K-means, k-medoids, and proposed DE models applied on the BRAC dataset. Number of clusters on the x-axis is shown and the y-axis has the DUNN index. The proposed model's DUNN index shows good compared to existing modes by varying the number of clusters. The dataset contains 57,063 genes information. The models cluster the genes causing the breast cancer data.

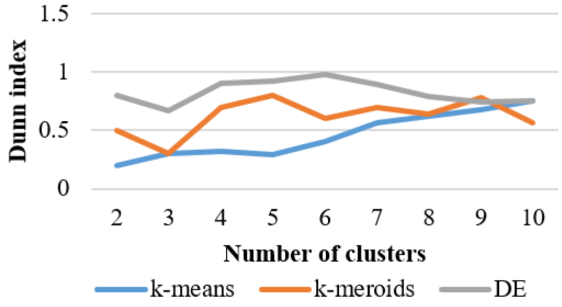

Figure 5. DUNN index on BRCA data

Here Figure 6 describes Silhouette index comparison analysis with K-means, k-medoids, and proposed DE models applied on the BRAC dataset. Number of clusters on the x-axis is shown and on the y-axis, it represents the Silhouette index. Silhouette index of the proposed model shows good compared to existing modes by varying the number of clusters. The dataset contains 57,063 genes information. The models cluster the genes that cause the breast cancer data.

Here Figure 7 shows the DB index comparison analysis with K-means, k-medoids, and proposed DE models applied on the BRAC dataset. Number of clusters on the x-axis is shown and the y-axis has the DB index. DB index of the proposed model shows good compared to existing modes by varying the number of clusters. The dataset contains 57,063 genes information. The models cluster the genes that cause the breast cancer data.

The three results Figures 5-7 are representing DUNN Index, Silhouette Index, DB Index on BRAC data respectively. Clusters are varying from two to ten clusters. By observing the clusters at sixth clusters we found better convergence in the DE model results. That is DUNN index is 0.98, Silhout index is 0.97 and DB Index is 0.08. These results indicated in METEBRIC data the most influenced gene has been found after making six clusters. Other models are failed to make proper convergence and results not promising like DE model.

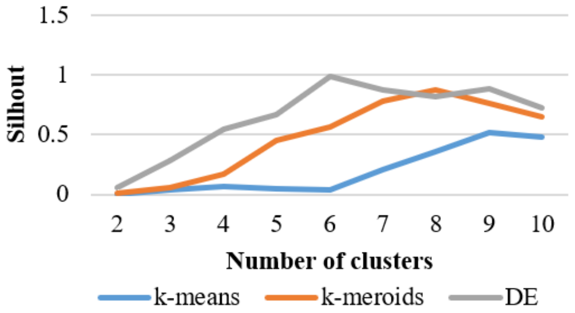

Figure 6. Silhouette index on BRCA data

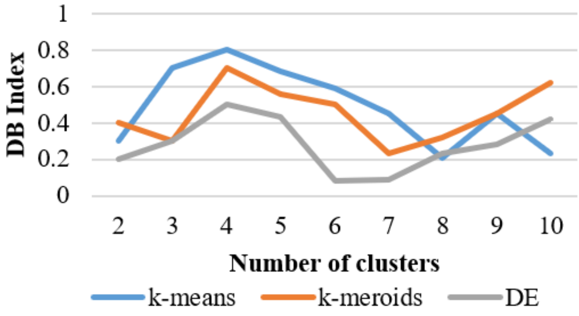

Figure 7. DB index on BRCA data

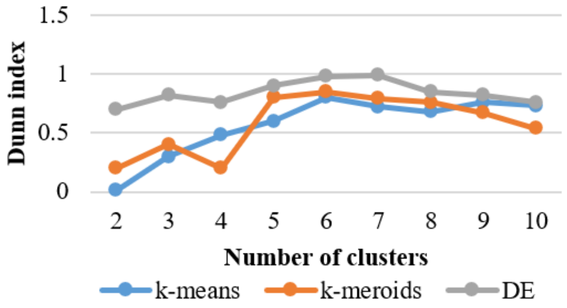

Figure 8. DUNN index on GWPS data

Here Figure 8 shows the DUNN index comparison analysis with K-means, k-medoids, and proposed DE models applied on the GWPS dataset. Number of clusters on the x-axis is shown and the y-axis has the DUNN index. DUNN index of the proposed model shows good compared to existing modes by varying the number of clusters. The dataset contains 348 breast cancer samples. The models cluster the genes that cause the breast cancer data.

Here Figure 9 shows the Silhouette index comparison analysis with K-means, k-medoids, and proposed DE models applied on the GWPS dataset. Number of clusters on the x-axis is shown and the y-axis has the Silhouette index. Silhouette index of the proposed model shows good compared to existing modes by varying the number of clusters. The dataset contains 348 breast cancer samples. The models cluster the genes that cause the breast cancer data.

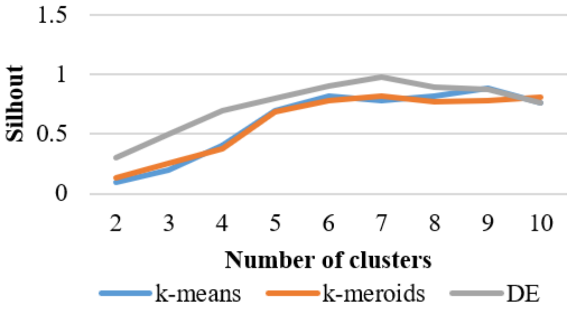

Figure 9. Silhouette index on GWPS data

Here Figure 10 shows the DUNN index comparison analysis with K-means, k-medoids, and proposed DE models applied on the GWPS dataset. Number of clusters on the x-axis is shown and the y-axis has the DUNN index. DUNN index of the proposed model shows good compared to existing modes by varying the number of clusters. The dataset contains 348 breast cancer samples. The models cluster the genes that cause the breast cancer data.

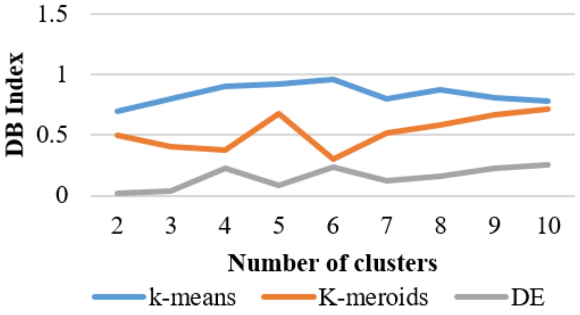

Figure 10. DB index on GWPS data

The three results Figures 8-10 are representing DUNN Index, Silhouette Index, DB Index on GWPS data respectively. Clusters are varying from two to ten clusters. By observing the clusters at seventh clusters we found better convergence in the DE model results. That is DUNN index is 0.99, Silhout index is 0.98 and DB Index is 0.1. These results indicated in GWPS data the most influenced gene has been found after making seventh clusters in GWPS dataset. Other models are failed to make proper convergence and results not promising like DE model.

5. CONCLUSIONS

High-dimensional gene expression data is difficult to cluster. Techniques for cluster analysis may miss an important partition even when a strong genetic signal is present. The researcher may not be aware of all the influences on the gene expression data, which might account for this discrepancy. As a result, generalising or making recommendations is difficult since the optimal clustering approach is reliant on the data itself. The genes related to breast cancer are discussed in this research, and the proposed DE model is applied to breast cancer cases. However, the findings imply that the properties of the data may impact the performance of the cluster. In this paper identify the most influenced gene from the gene expression data using DE clustering model. For that we identify the optimal number clusters can make most influenced gene. Here we take three different Brest cancer datasets for our experimentation. The fist dataset is METABRIC dataset and applied DE algorithm on after seven clusters the most influenced gene has been identified because all the experimental results are highly converged compared to other clusters. As same way BRPA and GWPS also made with DE model, BRCA results has been converged at six clusters and GWPS made in seven clusters. But the existing k-means and k-medoids models are failed to make convergence.

REFERENCES

[1] Brewer, H.R., Jones, M.E., Schoemaker, M.J., Ashworth, A., Swerdlow, A.J. (2017). Family history and risk of breast cancer: An analysis accounting for family structure. Breast Cancer Research and Treatment, 165(1): 193-200. https://doi.org/10.1007/s10549-017-4325-2

[2] Zahoor, J., Zafar, K. (2020). Classification of microarray gene expression data using an infiltration tactics optimization (ITO) algorithm. Genes, 11(7): 819. https://doi.org/10.3390/genes11070819

[3] Hambali, M.A., Oladele, T.O., Adewole, K.S. (2020). Microarray cancer feature selection: Review, challenges and research directions. International Journal of Cognitive Computing in Engineering, 1: 78-97. https://doi.org/10.1016/j.ijcce.2020.11.001

[4] Suresh, R.M., Dinakaran, K., Valarmathie, P. (2009). Model based modified k-means clustering for microarray data. In 2009 International Conference on Information Management and Engineering, Kuala Lumpur, Malaysia, pp. 271-273. https://doi.org/10.1109/ICIME.2009.53

[5] Maldonado, S., López, J. (2018). Dealing with high-dimensional class-imbalanced datasets: Embedded feature selection for SVM classification. Applied Soft Computing, 67: 94-105.

https://doi.org/10.1016/j.asoc.2018.02.051

[6] Dash, R. (2021). An adaptive harmony search approach for gene selection and classification of high dimensional medical data. Journal of King Saud University-Computer and Information Sciences, 33(2): 195-207. https://doi.org/10.1016/j.jksuci.2018.02.013

[7] Dashtban, M., Balafar, M. (2017). Gene selection for microarray cancer classification using a new evolutionary method employing artificial intelligence concepts. Genomics, 109(2): 91-107. https://doi.org/10.1016/j.ygeno.2017.01.004

[8] Chen, B., Tai, P.C., Harrison, R., Pan, Y. (2005). Novel hybrid hierarchical-K-means clustering method (HK-means) for microarray analysis. In 2005 IEEE Computational Systems Bioinformatics Conference-Workshops (CSBW'05), Stanford, CA, USA, pp. 105-108. https://doi.org/10.1109/CSBW.2005.98

[9] Uslan, V., Bucak, İ.Ö. (2010). Clustering-based spot segmentation of cDNA microarray images. In 2010 Annual International Conference of the IEEE Engineering in Medicine and Biology, Buenos Aires, Argentina, pp. 1828-1831. https://doi.org/10.1109/IEMBS.2010.5626430

[10] Wu, F.X., Zhang, W.J., Kusalik, A.J. (2003). Determination of the minimum sample size in microarray experiments to cluster genes using k-means clustering. In Third IEEE Symposium on Bioinformatics and Bioengineering, Bethesda, MD, USA, pp. 401-406. https://doi.org/10.1109/BIBE.2003.1188979

[11] Soares, C., Montgomery, L., Rouse, K., Gilbert, J.E. (2008). Automating microarray classification using general regression neural networks. In 2008 Seventh International Conference on Machine Learning and Applications, San Diego, CA, USA, pp. 508-513. https://doi.org/10.1109/ICMLA.2008.95

[12] Huang, H., Zhang, R., Xiong, F., Makedon, F., Shen, L., Hettleman, B., Pearlman, J. (2005). K-means+ method for improving gene selection for classification of microarray data. In 2005 IEEE Computational Systems Bioinformatics Conference-Workshops (CSBW'05), Stanford, CA, USA, pp. 110-111. https://doi.org/10.1109/CSBW.2005.82

[13] Karayianni, K.N., Spyrou, G.M., Nikita, K.S. (2012). Clustering microarray data using fuzzy clustering with viewpoints. In 2012 IEEE 12th International Conference on Bioinformatics & Bioengineering (BIBE), Larnaca, Cyprus, pp. 362-367. https://doi.org/10.1109/BIBE.2012.6399651

[14] Zhang, Y., Prinet, V., Wu, S. (2009). A principal component analysis based microarray data bi-clustering method. In 2009 2nd International Conference on Biomedical Engineering and Informatics, Tianjin, China, pp. 1-5. https://doi.org/10.1109/BMEI.2009.5305598

[15] Chen, T.S., Tsai, T.H., Chen, Y.T., et al. (2005). A combined K-means and hierarchical clustering method for improving the clustering efficiency of microarray. In 2005 International Symposium on Intelligent Signal Processing and Communication Systems, pp. 405-408. https://doi.org/10.1109/ISPACS.2005.1595432

[16] Guan, Y., Li, Y., Wang, Y., Zou, Y., Liu, M. (2010). The application of an improved K-means clustering method in microarray gene expressing data. In 2010 International Conference on Biomedical Engineering and Computer Science, Wuhan, China, pp. 1-4. https://doi.org/10.1109/ICBECS.2010.5462409

[17] Liang, Y., Liao, B., Zhu, W. (2017). An improved binary differential evolution algorithm to infer tumor phylogenetic trees. BioMed Research International, 2017: 5482750. https://doi.org/10.1155/2017/5482750

[18] Hosseiny, S.M., Rahmani, A.I., Derakhshan, M., Fatahizadeh, R. (2021). An intrusion detection system: Using a grasshopper algorithm. Ingénierie des Systèmes d'Information, 26(2): 171-177. https://doi.org/10.18280/isi.260204