Variance Score and Pearson Similarity based Unsupervised Feature Selection (VPUFS) for Sample Clustering in Microarray Gene Expression Data

Sourav Dutta
Department of CSE, NSEC
Kolkata, India
sourav.dutta@nsec.ac.in

Soumik Banerjee
Department of CSE, NSEC
Kolkata, India
soumikbanerjee381@gmail.com

Arghya Nath
Department of CSE, NSEC
Kolkata, India
arghyanath.mtechnsec2023@nsec.ac.in

Ankita Ghosh
Department of CSE, NSEC
Kolkata, India
ankitaghosh.mtechnsec2023@nsec.ac.in

Chandra Das
Department of CSE, NSEC
Kolkata, India
chandra.das@nsec.ac.in

Shilpi Bose
Department of CSE, NSEC
Kolkata, India
shilpi.bose@nsec.ac.in

***Abstract*— The analysis of gene expression data plays a pivotal role in cancer research, facilitating early detection, refined prognosis, and the development of targeted therapeutic strategies. Sample clustering from gene expression data provides a new way to perform cancer class discovery. However, the high dimensionality and complexity of gene expression datasets pose significant computational challenges, necessitating innovative approaches for feature selection and data reduction. This study proposes a novel framework called 'Variance Score and Pearson Similarity based Unsupervised Feature Selection (VPUFS)', an unsupervised feature selection technique to choose relevant features, to reduce high dimensionality, thereby helping as a preprocessing step to cluster cancer samples effectively. The unsupervised feature selection method leverages variance to calculate feature class relevance score, and similarity metrics to systematically identify and retain informative features while eliminating redundant ones. The proposed framework is rigorously evaluated across multiple microarray gene expression datasets, including Leukemia, Colon, Breast, and Prostate cancer. Comparative analyses demonstrate the superior performance of the proposed method in terms of classification and clustering accuracy and stability, over established unsupervised feature selection techniques.**

Keywords— Unsupervised feature selection, variance score, Pearson correlation, sample clustering, gene expression data.

I. INTRODUCTION

Cancer, a multifaceted and relentless disease, continues to exact a heavy toll on global health, standing as one of the leading causes of mortality worldwide. The magnitude of its impact is starkly highlighted by the World Health Organization's findings, which report a staggering 9.7 million deaths attributed to cancer in 2018 alone [1,2]. Tragically, this toll is projected to escalate drastically, with estimates suggesting an alarming increase in new cancer cases, potentially reaching a daunting 25 million annually by 2030 [1,2]. Despite remarkable strides in medical science, the prognosis for many cancer patients remains grim, underscoring the urgent need for innovative approaches to diagnosis, treatment, and management.

Research has previously shown that sufficient data from patient's clinical, environmental, and behavioural variables is not adequate for extremely precise cancer sample classification or prediction [3,4,5]. A number of genetic disorders with various biological characteristics have recently come to light as a result of various forms of bio-molecular data analysis. The study of these disorders are very useful for determining the response to various forms of treatment as well as for early cancer identification and prognosis [3,4,5].

In recent years, with the rapid evolution of genomic, proteomic, and high-throughput imaging technologies [6,7,8,9,10] there has been a surge in the accumulation of vast amounts of diverse biomolecular data from patients. Researchers are leveraging this wealth of information to pioneer advanced techniques aimed at early cancer detection, accurate prognosis, and enhanced therapeutic strategies to elevate patient survival rates. However, conventional laboratory-based approaches are proving inadequate due to the sheer volume and complexity of bio-molecular data, thus presenting formidable computational and analytical challenges, necessitating the development of innovative approaches to data analysis that can effectively navigate this intricate landscape. Consequently, computational methods such as statistical analysis, machine learning, and deep learning have emerged as indispensable tools in this realm [3,4,5].

The proliferation of computational and in-silico approaches has undoubtedly advanced cancer sample classification methodologies, primarily utilizing gene expression data. The aforementioned data facilitate various analyses, including but not limited to: (1) the development of classifiers for microarray sample classification using machine learning techniques, which may enhance the diagnosis of cancer patients and (2) the identification of genes with differential expression (different expression level in different sample classes) using supervised approach [2,11]. For first analysis a fundamental challenge persists due to the disproportionate ratio of sample size to gene count. Despite datasets often comprising hundreds of samples, they encompass thousands of genes, imposing a substantial computational burden on classification algorithms. Furthermore, amidst this vast gene pool, only a handful is informative, rendering classifier performance susceptible to noise. Consequently, the identification of informative biomarkers becomes paramount, not only for enhancing classification accuracy but also for elucidating cancer biology, evaluating disease risk, and devising targeted therapeutic interventions.

Apart from labelled datasets, in reality maximum microarray samples are available in which class labels are

missing. Proper grouping of cancer samples from these unlabelled datasets is a major challenge in cancer research. Herein, clustering cancer samples based on their gene expression profiles emerges as a pivotal strategy in cancer research [12,13,14]. Clustering techniques offer a complementary avenue to classification, facilitating the exploration of tumor molecular heterogeneity and the identification of novel subtypes. So, another typical and more exploratory study is to group patient samples (tissues) related to cancer using unsupervised approach [12,13,14]. The goal is to identify sets of samples with similar expression patterns, as this may help to identify novel cancer subtypes.

Moreover, clustering analysis unveils insights into the molecular mechanisms underpinning cancer progression, thus facilitating the identification of potential therapeutic targets and prognostic biomarkers. By bridging high-dimensional data analysis with actionable clinical insights, clustering emerges as a pivotal tool in cancer research, offering a nuanced understanding of tumor biology and guiding personalized treatment strategies. However, the application of clustering algorithms to high-dimensional gene expression data is fraught with challenges, including high dimensionality, noise, sparsity, and inherent biological variability. Traditional clustering algorithms, such as hierarchical clustering and k-means, often falter in the face of these complexities, yielding suboptimal results and limiting interpretability.

Hence reduction of high dimensionality of datasets is an important preprocessing step before cluster analysis [15,16].

Although several unsupervised dimensional reduction methods have been developed to reduce the complexity of gene expression data in cancer research, researchers continue to seek better alternatives for obtaining optimal results. In this regard, we have proposed a new framework called 'Variance Score and Pearson Similarity based Unsupervised Feature Selection (VPUFS)' for unsupervised genetic feature selection to reduce the dimensionality and noise of the dataset. The unsupervised feature selection method employs variance and pearson similarity metric to systematically identify and retain informative features while eliminating redundant ones. The method is implemented on various microarray datasets and predictive performance of the reduced dataset is evaluated with established models and contrasted with each other using different evaluation metrics.

II. PROPOSED WORK

The proposed algorithm 'Variance Score and Pearson Similarity based Unsupervised Feature Selection (VPUFS)' for dataset reduction utilizes feature score leveraging variance [85], and Pearson Correlation Coefficient as similarity metrics to systematically identify and retain features with unique and discriminative information while eliminating redundant features. By quantifying the relevance and distinguishing capabilities of individual features, and assessing the redundancy between features based on pearson similarity metrics, the algorithm constructs similarity matrices to identify clusters of redundant features. Subsequently, redundant features are systematically removed from the dataset, while representative features with high variance scores are retained. This streamlined feature selection process aims to enhance the efficiency and effectiveness of subsequent data analysis and modeling tasks by mitigating the risk of multi-collinearity and improving the interpretability and generalizability of predictive models.

The dataset reduction algorithm works by constructing a similarity matrix to identify and remove redundant features by computing pairwise pearson similarity between all features (genes) in the high-dimensional gene expression dataset.

Before data reduction some preliminary preprocessing techniques are applied on the dataset. These are discussed below.

A. Data Preprocessing:

The dataset is checked for any NULL values and corresponding rows and columns are deleted to get the complete dataset. Similar checking for duplicate rows and columns are also performed and removed accordingly to avoid the multi-collinearity issues.

After completing the preprocessing steps, we will retain our updated dataset containing headers and the target variable. Subsequently, we will remove the target column from the dataset and store it in another variable for further analysis.

B. Dataset Reduction Using Feature Score

To reduce the feature dimension, we are using variance score as class relevance score and Pearson similarity to measure the similarity for any given pair of features. Based on Pearson similarity we have calculated feature non-redundant score for every feature. Then we finally calculate final feature score for every feature.

Preliminaries:

Let the dataset be represented by a data matrix $D(m \times n) = d_{ki}$ where m is the no of samples and n is the no of features/attributes/genes in the dataset. The samples are represented by $E = \{E_1, E_2, E_3, \dots\dots\dots, E_m\}$ and the attributes/features/genes are represented by $A = \{A_1, A_2, A_3, \dots\dots, A_n\}$. Each sample is a n-dimensional attribute vector which contains n no of attributes and similarly each attribute contains m no of samples i.e. each feature is a m-dimensional sample vector. S is the reduced feature set.

1) Variance as Class Relevance Score

Variance, denoted by $\sigma^2(A_i)$ or $var(A_i)$, measures the spread or dispersion of data points around the mean within a dataset. It is calculated as:

$$var(A_i) = \frac{1}{m}\sum_{k=1}^{m}(d_{ki} - \mu_i)^2 \text{ ------(1)}$$

Where:

- (d_{ki}) represents k^{th} value of i^{th} feature.
- μ_i denotes the mean (average) of the i^{th} feature

High variance indicates that the data points are widely dispersed around the mean, reflecting significant fluctuations or variability within the dataset. This suggests a diverse range of values and captures important patterns or trends present in the data. On the other hand, low variance implies minimal variability, with data points clustered closely around the mean. This may indicate uniformity or a lack of significant variation within the dataset.

In the context of feature selection, features with high variance are considered valuable as they contain substantial information content and contribute significantly to the modeling process.

2) Pearson Similarity Coefficient for calculation of similarity among any pair of features

The Pearson correlation coefficient quantifies the strength of a linear relationship between two variables [17]. In essence, it seeks to establish a line that best represents the data. Pearson Correlation is given by:

$$P(A_i, A_j) = \frac{\sum_{k=1}^{m}((d_{ki} - \mu_i)(d_{kj} - \mu_j))}{\sqrt{\sum_{k=1}^{m}(d_{ki} - \mu_i)^2}\sqrt{\sum_{k=1}^{m}(d_{kj} - \mu_j)^2}} - - (2)$$

Where,

- (d_{ki}) represents k^{th} value of i^{th} feature.
- μ_i denotes the mean (average) of the i^{th} feature
- (d_{kj}) represents k^{th} value of j^{th} feature .
- μ_j denotes the mean (average) of the j^{th} feature

The Pearson correlation coefficient spans from +1 to -1. A score of 0 signifies no correlation between the two variables. A positive value suggests a positive association: as one variable increases, so does the other. Conversely, a negative value indicates a negative association: an increase in one variable corresponds to a decrease in the other.

3) Non-redundant score

The non-redundant score is designed to measure the uniqueness of a feature in relation to all other features in the dataset. This score helps in identifying features that provide unique information, minimizing redundancy among the selected features. The score is calculated as:

$$NoN\_redun\_score(A_i) = max\left(1 - P(A_i, A_j)\right) \; --(3)$$
$$For\ all\ j = 1\ to\ n\ and\ j \neq i$$

C. Proposed Final Feature score

The final feature score (FS) is a metric to be used to evaluate the overall importance of a feature by combining its individual variance with its non-redundancy score. This score aims to balance the feature's ability to capture significant information (variance) with its uniqueness relative to other features (non-redundancy). It is calculated as:

$$FS(A_i) = var(A_i) \times NoN\_redun\_score(A_i) --(4)$$

D. The proposed feature selection method

The steps of the proposed feature selection method named VPUFS is given below:

Algorithm: VPUFS

Input: Gene expression data matrix $D(m \times n) = d_{ki}$ where m is the no of samples and n is the no of features. The samples are represented by $E = \{E_1, E_2, E_3, \dots\dots\dots, E_m\}$ and the features are represented by $A = \{A_1, A_2, A_3, \dots\dots\dots, A_n\}$.

Output: The reduced feature set S.

1. Initialize $S = \emptyset$
2. For $i = 1\ to\ n$ do
   - 2.1 Calculate the variance score $var(A_i)$ from the feature set A using Equation (1).
   - 2. 2 For $j = 1\ to\ n\ and\ j \neq i$
     - 2.2.1 Calculate feature similarity of A_i using Pearson Correlation $P(A_i, A_j)$ sing equation (2).
   - 2.3 Calculate $NoN\_redun\_score(A_i)$ of feature A_i using equation (3).
   - 2.4 Calculate feature score $FS(A_i)$ of A_i using Equation (4).
3. Sort features in descending order according to their scores;
4. Select top q features to comprise the feature subset S.
5. End

After reducing the dimensionality of dataset, we will apply different clustering algorithms to cluster the cancer samples.

III. RESULT SECTION

In this research work, the performance of Proposed Algorithm is judged and compared with that of existing unsupervised feature selection algorithms like Laplacian Score (LS), Multi-Cluster Feature Selection (MCFS), Joint Embedding Learning and Sparse Regression (JELSR), Non-negative Discriminative Feature Selection (NDFS), and Local Discriminative Feature Selection (LDFS) on four different microarray gene expression data sets as given in table 1 [18,19,20,21,15].

Table1: Diseased Sample-based Dataset Description

Dataset Name	Species	Number of Columns/Genes	Number of Rows/Samples
Leukemia [22]	Human	7070	72
Colon [22]	Human	2000	62
Prostate [22]	Human	12600	136
Breast [22]	Human	7130	49

A. Description of evaluation indices

1) Cross-Validation Techniques

Cross-validation is a fundamental technique for assessing how well a predictive model generalizes to an independent data set. It is particularly useful in scenarios where the dataset is limited in size. Here, two widely used cross-validation methods: Leave-One-Out Cross-Validation (LOOCV), 5-Fold Cross-Validation and 10-Fold Cross-Validation have been used [22].

2) External Evaluation Metrics

a) Rand Index (RI)

The Rand Index (RI) is a metric used to evaluate the similarity between two data clustering, providing a straightforward measure of clustering accuracy. It is particularly useful in assessing how well the predicted clusters match the true class labels [12]. The RI value ranges from 0 to 1, with 0 indicating no agreement between the clusterings and 1 indicating perfect agreement. The RI is calculated as follows:

$$RI = \frac{TP + TN}{TP + TN + FP + FN} - - - - - (5)$$

where:

- **TP (True Positives)**: Pairs of elements that are in the same cluster in both the predicted and true clustering.

- **TN (True Negatives)**: Pairs of elements that are in different clusters in both the predicted and true clustering.
- **FP (False Positives)**: Pairs of elements that are in the same cluster in the predicted clustering but in different clusters in the true clustering.
- **FN (False Negatives)**: Pairs of elements that are in different clusters in the predicted clustering but in the same cluster in the true clustering.

b) Adjusted Rand Index (ARI)

While the Rand Index is a useful measure, it does not account for the possibility of agreement occurring by chance [12]. This limitation necessitates the use of the Adjusted Rand Index (ARI), which adjusts the RI for the chance grouping of elements. The ARI ranges from -1 to 1, where 1 indicates perfect agreement, 0 indicates random agreement, and negative values indicate less agreement than expected by chance. The ARI can be computed as:

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}\right] / \binom{n}{2}} \quad (6)$$

where:

- n_{ij} is the number of elements in the intersection between cluster i in the predicted clustering and cluster j in the true clustering.
- a_i is the number of elements in cluster i in the predicted clustering.
- b_j is the number of elements in cluster j in the true clustering.
- n is the total number of elements.

B. Evaluation of unsupervised feature selection

1) Comparative Analysis of VPUFS in terms of Sample Classification accuracy

The performance of the proposed unsupervised feature selection method has been compared with the existing unsupervised algorithms and full gene set in terms of sample classification accuracy using the LOOCV and 5-fold cross-validation methods for SVM classifier as shown in table 2 and table 3. VPUFS shows superiority for all datasets including the full gene set for the optimum no. of selected features. Table 4 shows the 10- Fold result. Though for the prostate cancer dataset, LDFS produces a slightly better accuracy that the proposed method, in other all cases performs better.

Table 2: Comparative Analysis in terms of LOOCV Accuracy

Dataset	VPUFS	Laplacian score	MCFS	JELSR	NDFS	LDFS	Full Gene Set
Colon	90.2	61.2	59.6	59.7	61.3	62.9	82.3
Breast	92.6	61.7	58.9	60.2	60.2	64.4	91.8
Prostate	93	67.7	82.5	80.6	79	88.2	91.9
Leukemia	98.7	65.2	63.8	65.2	59.7	63.8	98.6

Table 3: Comparative Analysis in terms of 5- fold Accuracy

Dataset	VPUFS	Laplacian score	MCFS	JELSR	NDFS	LDFS	Full Gene Set
Colon	86.9	61	55.7	58.2	58.2	64.3	76.2
Breast	91.8	61.2	56.4	57.1	57.1	62.2	82.4
Prostate	86.7	66.9	80.9	79.4	77.9	86	77.6
Leukemia	97.2	64.8	63.2	64.8	59.2	63.2	95.6

Table 4: Comparative Analysis in terms of 5- fold Accuracy

Dataset	VPUFS	Laplacian score	MCFS	JELSR	NDFS	LDFS	Full Gene Set
Colon	85.7	61.3	59.7	59.7	61.3	62.9	71.7
Breast	89	61.5	56.7	57.5	57.5	62.6	76.3
Prostate	86.1	67.7	82.3	80.6	79	87.1	69.8
Leukemia	95.7	65.3	63.9	65.3	59.7	63.9	91.4

2) Selection of parameter q as the no. of selected features

In order to decide the parameter q as the no. of top ranked features to be selected, the proposed VPUFS algorithm has been tested for different no. of top ranked features in terms of sample classification accuracy using LOOCV for the SVM classifier. The figure 1 shows the graph where the optimum value of q mostly in between 70 and 100 for the Colon, Breast and Prostate dataset while for Leukemia it is under 50.

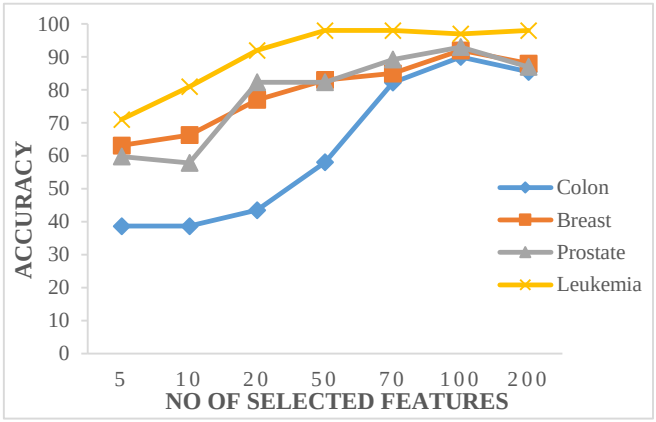

Figure 1: Accuracy for different no. of selected top ranked features to decide the optimum value for q.

3) Evaluation of different Sample clustering methods

In this section, we focus on evaluating the performance of various sample clustering methods applied to the reduced dataset produced by the proposed VPUFS. By employing RI and ARI as evaluation indices, we aim to provide an objective comparison of the clustering results obtained from different techniques. Our goal is to identify the most effective clustering method for each dataset, to accurately partition the data into meaningful clusters. Table 5 and table 6 show the rand index (RI) and adjusted rand index (ARI) of VPUFS for five sample clustering methods (K-means, Agglomerative, GMM, Spectral and SOM) when applied on the reduced dataset as well as the full dataset. The result shows the efficiency of the VPUFS method in almost every cases.

Table 5 : RI of k-means, Agglomerative, GMM, Spectral, and SOM on four gene expression datasets

	K-means		Agglomerative		GMM		Spectral		SOM	
Dataset	**Reduced**	**Full**	**Reduced**	**Full**	**Reduced**	**Full**	**Reduced**	**Full**	**Reduced**	**Full**
Colon	0.55	0.51	0.52	0.5	0.55	0.49	0.54	0.52	0.54	0.52
Breast	0.53	0.54	0.56	0.57	0.55	0.51	0.56	0.53	0.55	0.51
Prostate	0.51	0.5	0.5	0.51	0.51	0.5	0.53	0.54	0.52	0.5
Leukemia	0.65	0.53	0.64	0.59	0.66	0.57	0.56	0.53	0.61	0.58

Table 6 : ARI of k-means, Agglomerative, GMM, Spectral, and SOM on four gene expression datasets

	K-means		Agglomerative		GMM		Spectral		SOM	
Dataset	**Reduced**	**Full**	**Reduced**	**Full**	**Reduced**	**Full**	**Reduced**	**Full**	**Reduced**	**Full**
Colon	0.12	0.09	0.07	0.05	0.14	0.09	0.13	0.09	0.01	0
Breast	0.05	0.1	0.11	0.11	0.09	0.07	0.1	0.04	0.07	0.05
Prostate	0.02	0.01	0	0.01	0.02	0.01	0.06	0.1	0.04	0.01
Leukemia	0.32	0.17	0.3	0.2	0.35	0.2	0.16	0.11	0.37	0.25

IV. CONCLUSION

Cancer, a complex and relentless disease, continues to pose significant challenges to global healthcare systems. Despite remarkable advancements in medical research, the prognosis for many cancer patients remains bleak, highlighting the urgent need for innovative approaches to diagnosis, treatment, and management. This study aimed to address the computational challenges associated with analyzing high-dimensional gene expression data, a crucial aspect of cancer research. The performance of the proposed VPUFS algorithm represents the efficacy of the method for the dataset in absence of the sample class label information.

REFERENCES

[1] WHO. (2018). Cancer facts sheet. WHO http://www.who.int/mediacentre/factsheets/fs297/ en/

[2] Moshood A. Hambali, Tinuke O. Oladele, Kayode S. Adewole, "Microarray cancer feature selection: Review, challenges and research directions", International Journal of Cognitive Computing in Engineering, 2020, p. 78-97.

[3] Kourou, K., Exarchos, T. P., Exarchos, K. P., Karamouzis, M. V., & Fotiadis, D. I., "Machine learning applications in cancer prognosis and prediction", Computational and Structural Biotechnology Journal, 2015, 13, 8-17.

[4] Hambali, M. A., Oladele, O. A., & Adewole, K. Application of machine learning techniques in cancer prediction and treatment. SN Computer Science, 2020, 1(6), 1-12.

[5] Tabares-Soto R, Orozco-Arias S, Romero-Cano V, Segovia Bucheli V, Rodríguez-Sotelo JL, Jiménez-Varón CF, "A comparative study of machine learning and deep learning algorithms to classify cancer types based on microarray gene expression data", PeerJ Computer Science, 2020, 6:e270.

[6] Liu J, Cheng Y, Wang X, Zhang L, Wang ZJ., "cancer characteristic gene selection via sample learning based on deep sparse filtering. scientific reports", Nature, 2018, .8:8270.

[7] Su, A. I., Cooke, M. P., Ching, K. A., Hakak, Y., Walker, J. R., Wiltshire, T., ... & Hogenesch, J. B. (2001). Molecular classification of human carcinomas by use of gene expression signatures. Cancer research, 61(20), 7388-7393.

[8] Swan, A. L., Mobasheri, A., Allaway, D., Liddell, S., & Bacardit, J., "Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology. Omics: a journal of integrative biology, 2013, 17(12), 595-610.

[9] Crameri R, Schulz-Knappe P, Zucht HD. The future of post-genomic biology at the proteomic level: an outlook. Comb. Chem. High Throughput Screen. 2005;8(8):807–10.

[10] Pilling MJ, Henderson A, Gardner P., "Quantum cascade laser spectral histopathology: breast cancer diagnostics using high throughput chemical imaging", Analytical Chemistry 2017, 89(14):7348–7355 DOI 10.1021/acs.analchem.7b00426.

[11] Nagi S, Bhattacharyya DK.,"Classification of microarray cancer data using ensemble approach", Network Modeling Analysis in Health Informatics and Bioinformatics 2013, 2(3):159–173 DOI 10.1007/s13721-013-0034-x.

[12] Xianxue Yu, Guoxian Yu, Jun Wang, Clustering cancer gene expression data by projective clustering ensemble", PLOS ONE, 2017.

[13] J.P. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," J. Computational and Applied Math., vol. 20, no. 1, pp. 53-65, 1987.

[14] U. Alon, N. Barkai, D. A. Notterman, K. Gish, S. Ybarra, D. Mack, and A. J. Levine, "Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by Oligonucleotide arrays," Proc. Natl. Acad. Sci. USA, vol. 96, no. 12, pp. 6745–6750, 1999.

[15] Xiucai Ye et al., "Unsupervised Feature Selection for Microarray Gene Expression Data Based on Discriminative Structure Learning", Journal of Universal Computer Science, vol. 24, no. 6 (2018), 725-741 submitted: 22/5/17, accepted: 15/10/17, appeared: 28/6/18 © J.UCS

[16] Zhang, S., Wong, HS., Shen, Y., and Xie, D.: "A New Unsupervised Feature Ranking Method for Gene Expression Data Based on Consensus Affinity"; IEEE/ACM Transactions on Computational Biology and Bioinformatics. 9, 4 (July, 2012), 1257-1263

[17] Benesty, J., Chen, J., Huang, Y., Cohen, I. , "Pearson Correlation Coefficient. In: Noise Reduction in Speech Processing", Springer Topics in Signal Processing, 2009, vol 2. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-00296-0_5.

[18] He, X., Cai, D., and Niyogi, P.: "Laplacian score for feature selection"; Proc. Advances in Neural Information Processing Systems. 507-514.

[19] Multi-Cluster Feature Selection (MCFS) - "Unsupervised Feature Selection Using Multi-Cluster Feature Selection" by J. Li et al. (2011)

[20] Joint Embedding Learning and Sparse Regression (JELSR) - "Joint Embedding Learning and Sparse Regression for Unsupervised Feature Selection" by W. Liu et al. (2015)

[21] Nonnegative Discriminative Feature Selection (NDFS) - "Nonnegative Discriminative Feature Selection" by Z. Li et al. (2012)

[22] Local Discriminative Feature Selection (LDFS) - "Local Discriminative Feature Selection" by Y. Wang et al. (2015)