Performance Analysis of Dimensionality Reduction Techniques in Cancer Detection using Microarray Data

Swati B. Bhonde
Research Scholar,
Smt. Kashibai Navale College of Engineering,
Pune, India
swati.bhonde@gmail.com

Dr. Jayashree R. Prasad
Professor,
Sinhgad College of Engineering, Vadgaon,
Pune, India
jrprasad.scoe@sinhgad.edu

***Abstract*— Cancer is one of the major causes of deaths worldwide. This disease is more ghastly as it doesn't announce itself until it reaches in an advance stage. Still, mortality rate for cancer can be decreased if we diagnose & provide treatment at earliest. Though there are traditional clinical trials to predict a cancer there does not a single test which can correctly identify this disease. In the recent years DNA Microarray technology has been significantly used to analyze & predict the cancer. Analysis of gene expressions is not only interesting but also challenging as it is not only the concern of accuracy but also matter of life or death of a patient. DNA Microarray data is high dimensional, noisy & redundant, it makes task of classification more complicated as high computational cost is involved. Therefore feature selection & feature reduction becomes important task prior to classification. This paper presents comparative performance analysis of different dimensionality reduction techniques implemented on TCGA PANCANCER dataset.**

Keywords*— Cancer prediction, deep learning, dimensionality reduction, precision medicine, gene expressions

I. INTRODUCTION

According to the International Agency for Research on Cancer (IARC), 19.3 million new cases are registered for cancer and 10 million cancer deaths occurred in 2020 worldwide.[1] To handle this rising concern we need to adapt techniques like proteomic & genomic which gives deeper insight of gene expression of human being & detects specific biomarkers responsible for initiating particular disease. Now-a-days advent in Microarray technology has been widely used for cancer prognosis & diagnosis. It is used for concurrent examination of hundreds of genes activities in a single experiment. Microarray technologies have produced a huge amount of cancer genomic data publicly available which elevates an idea of the identification of candidate genes contributing to uncontrolled cell growth resulting in cancer. Analyzing gene expression data is crucial to find out harmful mutations and to avoid further consequences.[2] In recent years enormous research has been carried out in detecting type of cancer using genomic profiles. Still to achieve satisfactory cancer classification accuracy with the complete set of genes remains a great challenge, due to the high dimensions, small sample size, and presence of noise in gene expression data. To deal with this curse of dimensionality different algorithms are used by researchers.

As per the survey done in [2] gene expression dataset has open research issues like numbers of gene/attribute columns are greater than sample size, missing values, class imbalanceness, complex and noisy data thereby we need to have effective reduction technique to identify association amongst different genes. Because of these issues visualization & modeling of gene expression dataset becomes vital to discover important genetic aspect of a patient. Advanced machine learning & deep learning algorithms can help in this regard to deal with the curse of dimensionality. To reveal cause of cancer & to propose diagnostic method, special framework is proposed in [3] which can be later converted into lower dimension to save computational task of classification. Good dimensionality reduction algorithm enhances understanding, visualization & preserves most important characteristics present in original high dimensional dataset. As per the literature, two approaches are used for dimensionality reduction for gene expression dataset namely feature selection & feature extraction. [4] Statistical approaches like PCA, SOM, GA are investigated by researchers to find important genes for vaccine development.[5] Methods like co-relation and rank analysis are also used in some studies which reduces number of input variables Feature selection is a process of selecting appropriate attributes & eliminate unwanted one which helps in boosting performance of classifier & prevents model overfitting. Feature extraction is next phase which builds new feature subset from existing dataset. Filter methods, wrapper methods, embedded methods & hybrid approaches are used by researchers for feature selection.[6] Survey of all feature selection techniques with pros & cons is proposed in this paper which allows investigators to choose an `appropriate dimensionality reduction method.[7] Open research challenges associated with gene expression dataset are addressed using hybrid methods like maximum entropy covariance matrix & hybridized smoothed covariance estimators. Also flexibility & versatility of this method is compared using benchmark techniques to reduce dimensions & opt good accuracy for supervised classification.[8] To pick pertinent features from dataset optimized genetic algorithms are used on Malaria vector dataset which gave accuracy upto 85% using SVM algorithm. [4]. Additionally semantic web & data mining tools are used for combining data & experimental results from multiple sources.[9] Combination of consistency based subset evaluation and minimum redundancy maximum evaluation methods gives good classification performance accuracy. Here using PCA gives better accuracy over attribute selection method.[10] PCA is combined with SVM & Levenberg-Marquardt Backpropagation (LMBP) algorithm and it is concluded that SVM gives 94.98% accuracy & LMBP gives 96.07% accuracy.[11] It is further concluded that type of kernel & number of neurons proved as influential parameters during training process.

To solve the curse of dimensionality statistical methods

like regression & non-parametric regression methods are used on Leukemia Microarray dataset resulted in better accuracy.[12] Grouping Genetic Algorithm (GGA) implemented on RNA-seq data for five types of cancers gave average accuracy of 98.81 & standard deviation of 0.0174. It is further advised that parallel execution of GGA on several computers will reduce amount of time required significantly.[13] In order to select determinant genes from target dataset new feature selection methods like SVM based on recursive feature elimination and particle swarm optimization. Further optimal features selected are used for SVM classification. This algorithm has proved better in terms of accuracy, running time and extracting more prominent features from the dataset.[14] To rank important features as per their importance other methods like information gain, chi-square and absolute shrinkage & selection operations are also used to get optimized results.[15] Remaining part of paper is organized as follows: section-II discusses materials and methods for experimental design. Performance analysis of dimensionality reduction algorithms - PCA, t-SNE & UMAP is presented in Section-III followed by concluding remark on performance analysis of all algorithms is discussed in section-IV.

II. MATERIALS AND METHODS

A. TCGA PANCANCER Dataset:

To carry out this experiment we downloaded TCGA RNA PANCANCER dataset from UCI's repository.[16] Here, RNA-Seq gene expression levels measured by Illumina HiSeq platform. There are 801 sample & 20531 genes/features/attributes. Genes are identified with label gene_0 to gene_20530 This dataset covers following types of cancer diseases category –

BRCA - Breast carcinoma COAD - Colon adenocarcinoma

KIRC - Kidney Renal clear-cell carcinoma LUAD - Lung Adenocarcinoma

PRAD – Prostate Adenocarcinoma

Following figure shows number of samples used for each type of tumor.

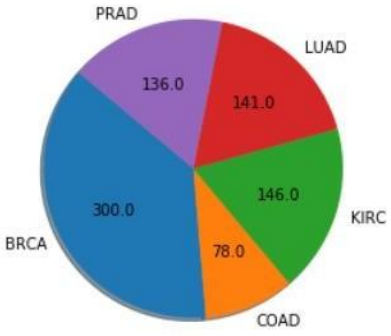

Fig. 1. Number of samples for each type

Above figure shows that dataset is not balanced because it has different number of samples for all types. BRCA group is having largest number of samples i.e. 300 & lowest i.e. 78 in COAD group. Fewer sample size still remains as a challenges as specified in [2].

B. Gene Selection:

In Microarray cancer dataset instead of analyzing complete gene set, only prominent biomarker genes can be selected in order to save computational cost & enhance classifier accuracy. Following figure shows histogram plotted on our dataset.

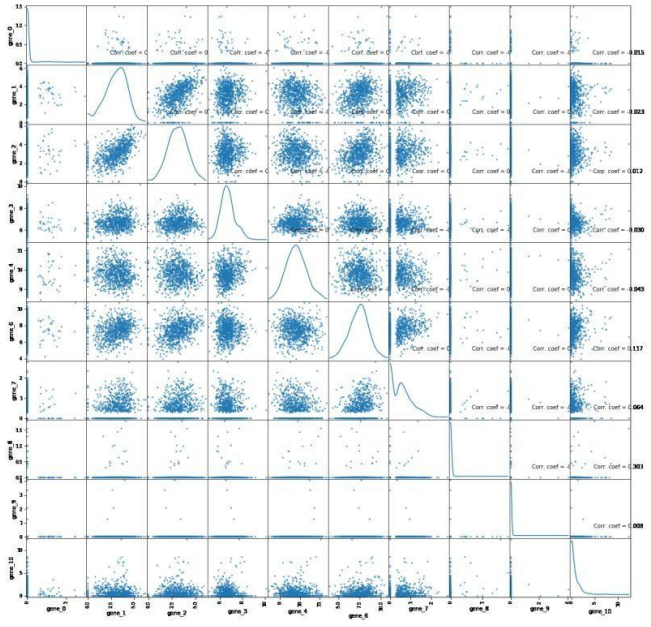

Fig. 2. Histogram plotted on dataset

Appropriate statistical test is used to rank the genes based on their importance and then biologically significant genes are selected in order to predict particular disease. Genes identified in this process are used for tumor analysis, drugs & vaccine development. Methodology used for this work is shown in figure -3.

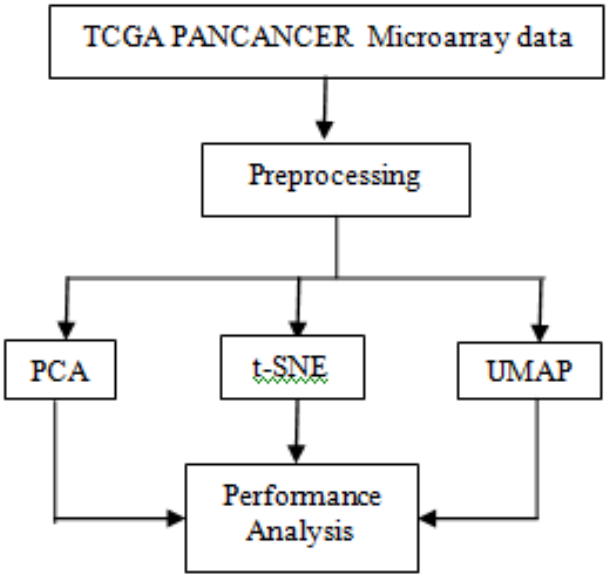

Fig. 3. System architecture

Our system has two main stages preprocessing & dimensionality reduction using three algorithms PCA, t- SNE & UMAP.

III. IMPLEMENTATION ANALYSIS

This section elaborates results of preprocessing & three dimensionality reduction algorithms.

A. Preprocessing phase:

Following table shows structure of records in our dataset. First of all, column_0 indicates record number which is not sort of any data or feature, so it can be eliminated. Next, we look for any entry with NULL value & omitted that as that feature also will not have any influence of prediction. Further, we found there are total 267 entries with same minimum & maximum value so deleted them as well.

TABLE I. STRUCTURE OF DATASET

	gene_0	gene_1	gene_2	gene_3	gene_4	gene_5	gene_
sample_0	0.0	2.017209	3.265527	5.478487	10.431999	0.0	7.17517
sample_1	0.0	0.592732	1.588421	7.586157	9.623011	0.0	6.81604
sample_2	0.0	3.511759	4.327199	6.881787	9.870730	0.0	6.97213
sample_3	0.0	3.663618	4.507649	6.659068	10.196184	0.0	7.84337
sample_4	0.0	2.655741	2.821547	6.539454	9.738265	0.0	6.56696

This dataset is further given to dimensionality reduction phase.

B. Dimensionality reduction techniques:

It is required, to remove redundancy and fetch irrelevant features by reducing the feature ratio of the samples which helps in decreasing the probability of overfitting. This section discusses & explores results obtained for PCA, t-SNE & UMAP algorithm.

Principal Component Analysis:

PCA is linear dimensionality technique & it preserves global structure of data. Applying dimensionality reduction with PCA will reduce dimensional complexity because the Microarray data will extract its features using eigenvectors and eigenvalues that have been obtained. Steps for dimensionality reduction algorithm using PCA[17], are described below:

1. Let X be an input matrix for PCA. X is training data composed of a n-vector with data dimension m.
2. Calculate the mean data of each dimension (X) using equation 1:

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \qquad (1)$$

Where: n = Number of samples or number of observation data, Xi = Observation data

3. Calculate the covariance matrix (C_X) using Equation (2):

$$C_x = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T \quad \ldots\ldots (2)$$

Where: n = Number of samples or number of observation data

Xi = Observation data X = mean data

4. Calculate the eigenvectors (v_m) and eigenvalues (λ_m) of the covariance matrix using Equation (3):

$$C_m v_m = \lambda_m v_m \quad \ldots \quad \ldots(3)$$

5. Sort the eigenvalues in descending order
6. Principal Component (PC) is a collection eigenvector corresponding to the sorted eigenvalues in step 5
7. PC dimension will be reduced based on the eigenvalues.

We executed PCA on our dataset which resulted in following output.

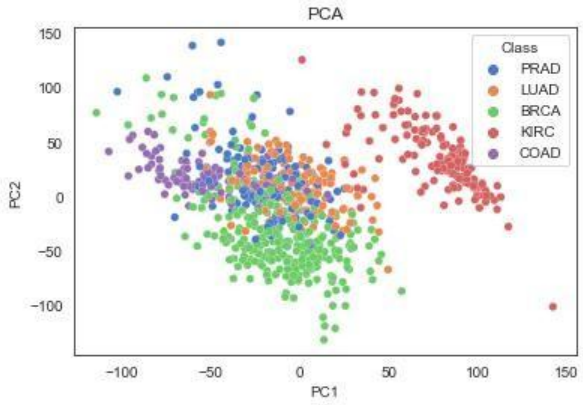

Fig. 4. Output of PCA

Average time required for executing PCA algorithm is 6.05 Seconds. This algorithm does not consider hyperparameters & is very sensitive to outliers.

t- Stochastic Neighbor Embedding:

It is unsupervised non-linear technique primarily used for visualizing high dimensional data. This algorithm tries to preserve clusters of data. Following are the steps for t-SNE algorithm:

1. Measure similarities between points in high-dimensionaldataset.
2. Find second set of probabilities using Cauchy distribution.
3. Measure the difference between probability distributionof 2-dimensional spaces.

We executed t-SNE algorithm on our dataset & we got following output.

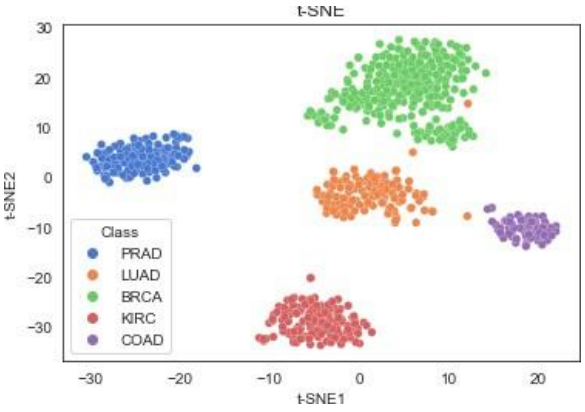

Fig. 5. Output of t-SNE

Time required to execute t-SNE algorithm is 29.82 Seconds. This algorithm handles outliers properly & also involves hyperparameters as perplexity, learning rate & number of steps.

IV. UNIFORM MANIFOLD APPROXIMATION AND PROJECTION:

This is new dimensionality reduction method introduced in year 2018. It can be directly applied to sparse data. UMAP is based on manifold learning technique & ideas from topological data analysis. Following are steps for implementing UMAP algorithm:

1. Compute a graphical representation of a dataset - fuzzysimplical complex.
2. Through stochastic gradient descent, optimize a low-dimensional embedding of the graph. Here, we use deep neural network that learns a parametric relationship between data and embedding.

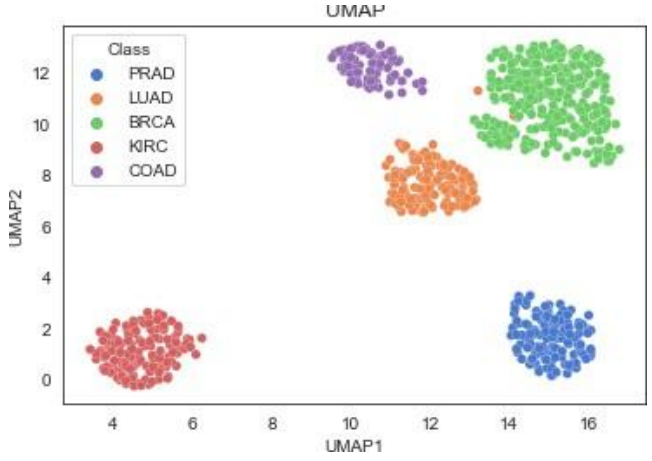

Fig. 6. Output of UMAP

Time required for UMAP algorithm us 7.10. UMAP too-much faster than t-SNE. UMAP maintains stochasticity even though it is initialized randomly with PCA.

V. PERFORMANCE ANALYSIS

Following table shows time required for executing each of these algorithms.

TABLE II. COMPARATIVE ANALYSIS OF DIMENSIONALITY REDUCTION TECHNIQUES

Algorithm	Time
Principal Component Analysis	6.05
t-Stochastic Neighbor Embedding	29.82
Uniform Manifold Approximation and Projection	7.10

Above table shows that t-SNE is slower than UMAP & PCA. Fastest algorithm is PCA & it is able to reconstruct the original data set. UMAP preserves pairwise Euclidean distances considerably better than tSNE. UMAP preserves the shapes of the clusters better than tSNE. PCA is unsupervised learning algorithm & it works by identifying the hyperplane which is closest to the data and then projects the data on that hyperplane while retaining most of the variation in the data set. As our dataset is already complex & noisy, to capture important insights from this multi-dimensional data, we used PCA. It makes the data more linearly separable within 500. When we executed PCA on our dataset, we got following result. components finally we get transformed features of genes 500 columns

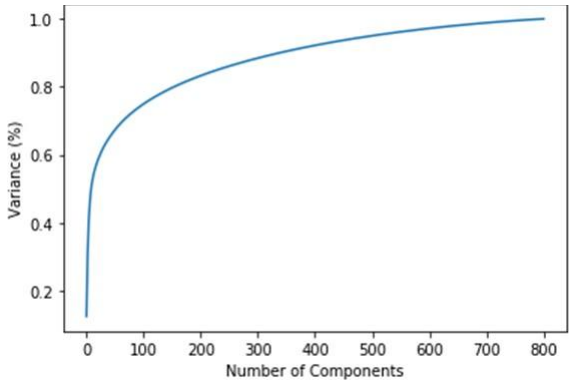

Fig. 7. PCA Variance graph

From the above plots we can interpret 500 is optimal components which captures most variance.

```
Performing Principal Component Analysis (PCA) ...
          PC1        PC2 Unnamed: 0 Class
0 -57.446987  95.410981   sample_0  PRAD
1 -16.919430   0.732470   sample_1  LUAD
2 -70.345218 -19.303327   sample_2  PRAD
3 -49.161591  -9.227586   sample_3  PRAD
4 -18.132534 -51.327797   sample_4  BRCA
```

Fig. 8. Principal Components

Above figure shows values for two principle components. Genes get compressed within 500 features components of PCA. Count of features extracted after PCA is 500

VI. CONCLUSION

Gene selection plays very important role in classification of cancer type using learning technique. And as gene expression data is very complex, noisy & redundant one, prominent gene selection remains as a challenge. This paper compares PCA, t-SNE & UMAP algorithm & discusses results obtained for each of them. It is observed that PCA comparatively takes less time & preserves original structure of data.

VII. ACKNOWLEDGEMENT

I am deeply indebted to Dr. Jayashree R. Prasad under whom I am working as a research scholar. She has been proved as a driving force for me to carry out this research work. Her patience motivation & immense knowledge has helped me a lot during this amazing journey. I express my sincere thanks towards her for believing in me & encouraging me to carry out this research work. Last but not least, I am obliged to my family & friends for supporting me emotionally & spiritually throughout my life.

REFERENCES

[1] Globocan, "New Global Cancer Data." https://www.uicc.org/news/globocan-2020-new-global-cancer-data.

[2] M. A. Hambali, T. O. Oladele, and K. S. Adewole, "Microarray cancer feature selection: Review, challenges and research directions," *Int. J. Cogn. Comput. Eng.*, vol. 1, no. October, pp. 78–97, 2020, doi: 10.1016/j.ijcce.2020.11.001.

[3] A. Bhola and S. Singh, "Visualisation and Modelling of High-Dimensional Cancerous Gene Expression Dataset,"*J. Inf. Knowl. Manag.*, vol. 18, no. 1, 2019, doi: 10.1142/S0219649219500011.

[4] M. O. Arowolo, M. O. Adebiyi, A. A. Adebiyi, and O. J. Okesola, "A Hybrid Heuristic Dimensionality Reduction Methods for Classifying Malaria Vector Gene Expression

Data," *IEEE Access*, vol. 8, pp. 182422–182430, 2020,doi: 10.1109/access.2020.3029234.

[5] C. S. Kong, J. Yu, F. C. Minion, and K. Rajan, "Identification of biologically significant genes from combinatorial microarray data," *ACS Comb. Sci.*, vol. 13, no. 5, pp. 562–571, 2011, doi: 10.1021/co200111u.

[6] S. Khalid, T. Khalil, and S. Nasreen, "A survey of feature selection and feature extraction techniques in machine learning," *Proc. 2014 Sci. Inf. Conf. SAI 2014*, pp. 372– 378, 2014, doi: 10.1109/SAI.2014.6918213.

[7] F. Song, D. Mei, and H. Li, "Feature selection based on linear discriminant analysis," *Proc. - 2010 Int. Conf. Intell. Syst. Des. Eng. Appl. ISDEA 2010*, vol. 1, pp. 746– 749, 2010, doi: 10.1109/ISDEA.2010.311.

[8] E. Pamukçu, H. Bozdogan, and S. Çalik, "A novel hybrid dimension reduction technique for undersized high dimensional gene expression data sets using information complexity criterion for cancer classification," *Comput. Math. Methods Med.*, vol. 2015, 2015, doi: 10.1155/2015/370640.

[9] F. Rafii, B. D. R. Hassani, and M. A. Kbir, "New approach for microarray data decision making with respect to multiple sources," *ACM Int. Conf. Proceeding Ser.*, vol. Part F1294, 2017, doi: 10.1145/3090354.3090463.

[10] J. Taveira De Souza, A. Carlos De Francisco, and D. C.De Macedo, "Dimensionality Reduction in Gene Expression Data Sets," *IEEE Access*, vol. 7, pp. 61136– 61144, 2019, doi: 10.1109/ACCESS.2019.2915519.

[11] Adiwijaya, U. N. Wisesty, E. Lisnawati, A. Aditsania, and D. S. Kusumo, "Dimensionality reduction using Principal Component Analysis for cancer detection based onmicroarray data classification," *J. Comput. Sci.*, vol. 14, no. 11, pp. 1521–1530, 2018, doi: 10.3844/jcssp.2018.1521.1530.

[12] A. Antoniadis, S. Lambert-Lacroix, and F. Leblanc, "Effective dimension reduction methods for tumor classification using gene expression data," *Bioinformatics*,vol. 19, no. 5, pp. 563–570, 2003, doi: 10.1093/bioinformatics/btg062.

[13] A. Lopez-Rincon, M. Martinez-Archundia, G. U. Martinez-Ruiz, A. Schoenhuth, and A. Tonda, "Automatic discovery of 100-miRNA signature for cancerclassification using ensemble feature selection," *BMC Bioinformatics*, vol. 20, no. 1, pp. 1–17, 2019, doi: 10.1186/s12859-019-3050-8.

[14] Y. Zhang, Q. Deng, W. Liang, and X. Zou, "An Efficient Feature Selection Strategy Based on Multiple Support Vector Machine Technology with Gene Expression Data," *Biomed Res. Int.*, vol. 2018, 2018, doi: 10.1155/2018/7538204.

[15] O. Rehman, H. Zhuang, A. M. Ali, A. Ibrahim, and Z. Li, "Validation of miRNAs as breast cancer biomarkers with a machine learning approach," *Cancers (Basel).*, vol. 11, no. 3, pp. 1–10, 2019, doi: 10.3390/cancers11030431.

[16] UCI, "TCGA Pancancer dataset," [Online]. Available: https://archive.ics.uci.edu/ml/datasets/gene+expression+cancer+RNA-Seq#.

[17] W. Astuti and Adiwijaya, "Support vector machine and principal component analysis for microarray data classification," *J. Phys. Conf. Ser.*, vol. 971, no. 1, 2018, doi: 10.1088/1742-6596/971/1/012003.