International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8, Issue-1, May 2019

Microarray Gene Expression Data and Performance Analysis of Various Missing Data Imputation Techniques

K Ishthaq Ahmad, Shaheda Akthar

Abstract*: Data with missing value is a curse for valuable data, especially in the case of microarray data analysis. Usually, Microarray gene expression data looks like matrix data with a set of genes under various environmental conditions. Microarray gene expression data further undergoes processing in terms of deviations and some other statistical measures. These statistical measures require microarray gene expression data with complete values, but in general condition, data consisting of a certain percentage of missing values. Results which obtain on these missing gene expression data are inconsistent and this result deviates from the original. So that it is necessary to impute the missing values before any estimation is done. In this paper, we have analyzed the performance of various imputation methods based on two real-time microarray datasets.

***Index Terms*: Missing data, Microarray array, Imputation, Predictive mean, Random-Forest.**

I. INTRODUCTION

Microarray gene expression data consisting of a huge number of genes under different experimental conditions. These Microarray gene expression data obtained through a systematic procedure of DNA synthesis. During the synthesis of DNA usually, they are taking one good sample and one contaminated sample. These two samples are amalgamated through a fluorescent liquid and placed on some glass place. The glass plate is thoroughly washed and the dust on the plate is cleaned and exposed to some light object. The image of the glass place is captured under the electronic microscope. Thus the image captured is processed through several image processing techniques to capture the intensity of fluorescent dots. These intensity values are normalized to get our gene expression data. The captured microarray data some time consists of missing values at their respective position of genes. The structure of microarray gene expression data similar to the matrix from where the rows represent genes and columns represents experimental conditions. This microarray gene expression data consists of lots of missing entries. These missing entries in the data are because of negligence during the extraction of microarray gene expression data. These missing entries are in the data because of improper washing and dust particles remain after washing on the glass plates. [9][10] Basically these missing data phenomenon fall under three categories missing completely at random (MCAR), missing at random (MAR) and missing not at random (MNAR). In the missing completely at random where the missing values in the data neither dependent on the pattern of observed nor pattern of unobserved values. Several proposed algorithms will fall into such categories. Missing at random (MAR) where the pattern of missing values depends on the observed pattern of missing data. Missing not at random (MNAR) in this pattern of missing data depends on the unobserved data pattern.

$$MicroArray\ data = D_{ij} = \begin{pmatrix} a_{11} & a_{12} & a_{13} & . & . & a_{1n} \\ a_{21} & a_{22} & a_{23} & . & . & a_{2n} \\ a_{31} & a_{32} & ? & . & . & a_{3n} \\ . & . & . & . & ? & . \\ . & . & . & . & . & . \\ a_{m1} & ? & ? & . & . & a_{mn} \end{pmatrix} \quad (1)$$

D_{ij} is the microarray data with m genes and n number of samples. From equation (1) a_{ij} represents the data value belongs to the i^{th} gene with j^{th} sample. From the data matrix, it is observed that some cells have entries and some does not have the values in it. Total data values can be divided as observed values (and unobserved values $a_{ij} = O_{ij}$ and unobserved values (Missing entries) ($a_{ij} = ?$) which is represented as ? in the matrix. From data matrix D_{ij} through analysis is made to identify what value can be placed in the missing entries replacing the NA with values. [1] In this paper author has discussed about the influence of various hierarchical clustering techniques on missing data estimation algorithms and also how this missing data values can affect the accuracy of clustering algorithms. [2] The author has imputed the missing values at higher dimensional space by using the orthogonal coding input scheme. [3] An integrated neighboring gene information has been incorporated in the place of missing values. [4] In this paper the author has extended the local least square method by extracting principal components from the datasets. [5] The author has proposed an hybrid approach of missing data estimation techniques by incorporating both local structure and global correlation to improve Bayesian principal component analysis and local least square estimation. [6] Regression based imputation has been improved by taking the correlation among the data and selecting the similar gene based on Pearson's correlation coefficient. [7]

Revised Manuscript Received on 30 May 2019.
***** Correspondence Author
K Ishthaq Ahmad*, Research Scholar, Dept. of Computer Science and Engineering, Acharya Nagarjuna University, Guntur.
Dr.Shaheda Akthar, Registrar$_{FAC}$ Dr.Abdul Haq Urdu University, Kurnool.

In this missing values are estimated based on two clusters, one performed to identify the similar genes and another performed column-wise to extract similar experimental conditions. [8] In this paper authors proposed an imputation method based on the difference of measured value during the cluster with original value and cluster with missing values, this difference is used for measuring the missing value through Gaussian mixture model. In this paper, we have investigated the performance analysis of various missing data imputation algorithms through two real-time datasets.

II. REVIEW OF VARIOUS DATA IMPUTATION TECHNIQUES

Missing data imputation algorithms broadly classified into three categories, one local based imputation algorithm which takes the local structure and the relation among the data. Another Global based missing data imputation algorithm in which the algorithm takes the external information or global information and incorporate this information into their respective algorithms. Last is a hybrid based algorithm which takes the information of both local and global. Some imputation methods are classified as general and advanced methods. Following are some imputation algorithms which falls under general and advanced methods.

A. List wise deletion

In this method, variables are eliminated from the complete dataset whose data values are missing. After the elimination of variables, the set becomes to get rid of incomplete values. Now we can make the analysis on the remaining complete data set [11].

B. Pair wise Deletion

In this case, missing data has been deleted by case by case analysis. This method is adopted under MCAR (missing completely at random) strategy [12].

C. Mean/Median/ Mode Imputation

In this method, missing values/entries are replaced by either row mean/median/mode or column mean/median/mode of the dataset. In this type of imputation mostly produces bias during the estimation of missing values. [11]

D. Decision Tree based Imputation [13]

The decision tree is one of the finest techniques of decision making and used for the classifying the items based on their respective computed properties. Generally, the dataset is arranged in a tree-based structured, where root nodes represent the observations or intermittent values and leaf nodes represent labels to which these data items belong. The construction of the decision tree is based on the splitting procedure in which the Gini index and the Split index values are computed. The main procedure of decision tree imputation is splitting the datasets into two sets, one training set with complete data values and another testing set with missing values. Through the method of classification and prediction, we can predict the missing values in the datasets.

E. K nearest Neighbor Imputation [14]

This imputation finds the nearest neighbor to the observed values and places them in the missing values, which are very nearest to the missing values. Nearest values are computed mostly based on the Euclidian distance between the values of the datasets.

K nearest neighbor is a classification algorithm, which identifies what group a data point belongs by identifying the nearest data points. The k-nearest-neighbor is an example of a "lazy learner" algorithm because it does not generate a model of the data set beforehand.

Input:

T{Training Data}

K{Number for classify}

x'{Input object to classify}

Output:

c{ Class to which x' is assigned}

$N \leftarrow 0$

for all $v \in T$ *do*

if $|N| \leq K$ *then*

$N \leftarrow N \cup \{v\}$

else if $\exists u \in N$ *such that* $d(x', u) \geq d(x', v)$ *then*

$N \leftarrow N - \{u\}$

$N \leftarrow N \cup \{v\}$

end if

end for

c= class to which the most u $\in N$ *are classified.*

F. Predictive mean matching (PMM) Imputation 15,16].

Predictive mean matching has been discussed several times in the history but its main use came to known recently. Usually, the PMM method has been used for single data imputation, but in recent days it extended to multiple missing values imputation. Let us consider there is some variable X whose some data values are missing and remaining complete variables are considered as Y.

Step 1. Find the cases in which no missing data in X estimate the linear regression of X on Y, to yield a set of coefficients called β.

Step 2. From the posterior distribution select the random values to produce a new set of coefficients called $\hat{\beta}$. From multivariate normal distribution estimate the mean β and estimate the covariance matrix of β.

Step 3. Using $\hat{\beta}$extract the predictive values of X for all cases (missing and non-missing data)

Step 4. Each case of missing values of X predict the nearest predicted values of X.

Step 5. For all nearest cases randomly select the values and substitute in place of missing value.

Step 6. Repeat the steps 2 to 5 for each completed dataset.

G. Bayesian Linear Regression Imputation [16][BLR]

In this instead of the point estimate, it considers as distribution. Suppose the response variable Y which is to be estimated from distribution rather than regression.

$$Y \sim N(\beta^T X, \sigma^2 I) \quad (2)$$

From equation 2, it is observed that out of Y is generated from the normal distribution. The mean values are computed from the product of coefficients of linear regression and matrix of values of X, the variance is calculated from the square of the standard distribution is multiplied by identity matrix.

Algorithm: *let us consider the Y as a dataset with missing values and X_{obs} as observed values with no missing entries.*

Step 1: *Calculate the cross product of matrix* X_{obs} $H=X'_{obs}X_{obs}$

Step 2: *Calculate* $V = (\sum_{obs} X_i' X_i)^{-1}$

Step 3: *From posteriori* σ^2 *is* $\sigma_1^2(n_1 - q)$ *divided by a* $\chi^2_{n_1-q}$ *random variable, and* β *given* σ^2 *is normal with mean* $\widehat{\beta_1}$ *and variance and covariance matrix* $\sigma^2 V$

Step 4: *Calculate regression weights* $\hat{\beta}_1 = V[\sum_{obs} X_i Y_i]$

Step 5: *Calculate* $\sigma_1^2 = \frac{\sum_{obs}(Y_i - X_i\widehat{\beta_1})^2}{n_1 - q}$

Step 6: *Draw q independent variants from N (0,1) from the vector* $\dot{z_1}$

Step 7: *Calculate* $\dot{\beta} = \widehat{\beta_1} + \dot{\sigma} V^{1/2} Z$

Step 8: *Draw* n_0 *independent from N (0,1) variants in* $\dot{z_2}$

Step 9: *Calculate* n_0 *value from* $Y_{mis} = X_{mis}\dot{\beta} + \dot{z_2}\,\dot{\sigma}$

H. Random-Forest Imputation

Usually works on mixed types of data sets either categorical or continuous data. The basic building block is the decision tree. The fundamental idea behind a random forest is to combine many decision trees into a single model. Individually, predictions made by decision trees (or humans) may not be accurate, but combined together; the predictions will be closer to the mark on average. It can handle non-linear relation and complex structure in the data sets. This algorithm is based on random forest [17] [Breiman 2001] Main advantage of this algorithm is it can run in parallel to save computational time.

III. Datasets used for Imputation techniques [18, 19]

Prostate cancer microarray dataset has been used for this experiment. The dataset consists of 2135 genes and 102 samples. For the computational simplicity, we reduced the number of the sample from 102 to 24 and genes remain the same. In Fig 1 the plots a, b, c, d shows the cluster plot of missing values, which shows the missing patterns from the samples as well as genes.

Second data set belongs to the CDC15 yeast gene expression data set of Spellman dataset. This dataset consists of 4381 genes and 24 samples. For our experiment, we have eliminated the first sample which is not a numeric quantity and the remaining 23 samples are used for our experiment. In Fig 2 the plots a, b, c, d shows the cluster plot of missing values, which shows the missing patterns from the samples as well as genes.

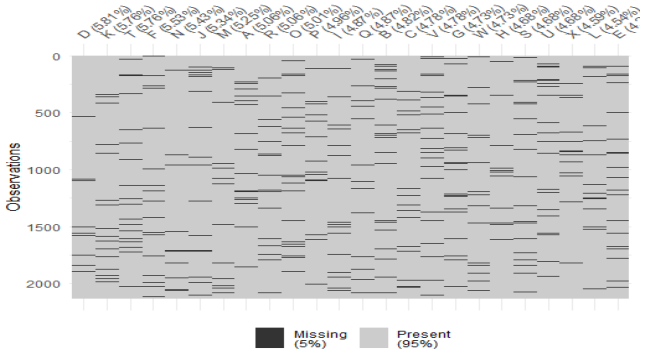
IV. Root mean square error (RMSE)[20]

The Root Mean Square Error (RMSE) (also called the root mean square deviation, RMSD) is frequently used to measure the difference between values predicted by a model and the values actually observed from the environment that is being modeled. These individual differences are also called residuals, and the RMSE serves to aggregate them into a single measure of predictive power.

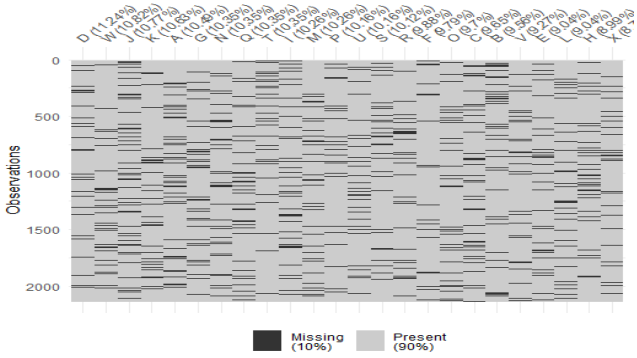
The RMSE of a model prediction with respect to the estimated variable X_{ij}^{esti} is defined as the square root of the mean squared error:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}\left(X_{ij}^{O} - X_{ij}^{esti}\right)^2}{mn}} \quad (3)$$

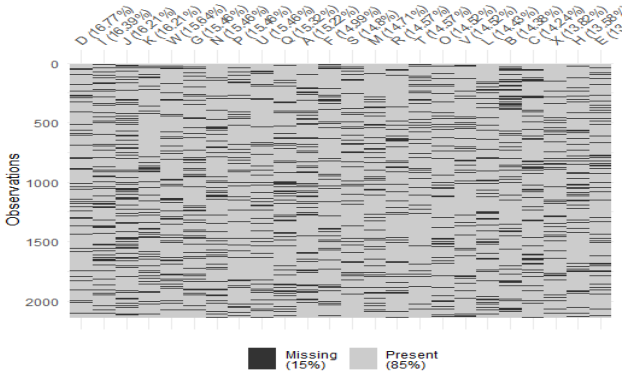
Where X_{ij}^{O} is observed values and X_{ij}^{esti} are estimated values of i[th] gene and j[th] experiment



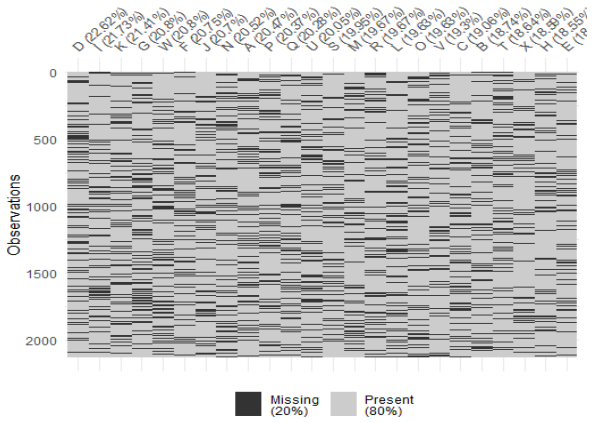
(a)



(b)



(c)



(d)

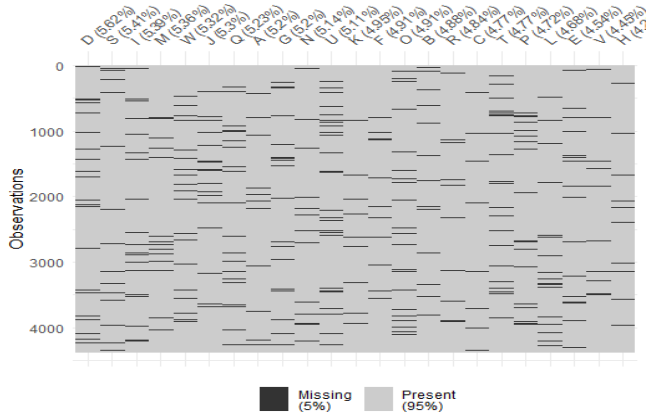
Figure. 1 Percentage of missing value pattern in the Prostate cancer data set a)5% b)10% c)15% d)20%

Table. 1 RMSE values estimated for different missing data imputation techniques under various percentage of missing data

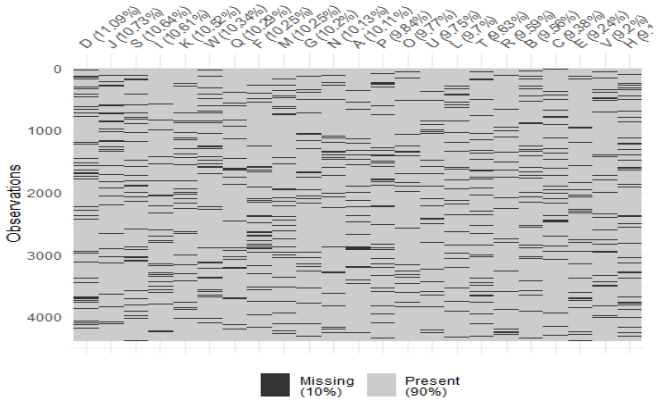
Dataset used: Prostate				
Percentage of missing data	5	10	15	20
Number of missing values	2562	5124	7686	10248
Number of complete values	48678	46116	43554	40992
Performance analysis of various missing data Imputation Technique (RMSE value)	Percentage of missing data			
	5	10	15	20
mean	0.009	0.006	0.005	0.004
median	0.009	0.006	0.005	0.004
mode	0.024	0.017	0.014	0.012
decision	0.015	0.015	0.016	0.026
KNN	0.015	0.013	0.011	0.011
PMM	0.015	0.012	0.011	0.009
BLR	0.019	0.015	0.011	0.01
R.FOREST	0.012	0.008	0.007	0.006

Table. 2 RMSE values estimated for different missing data imputation techniques under various percentage of missing data

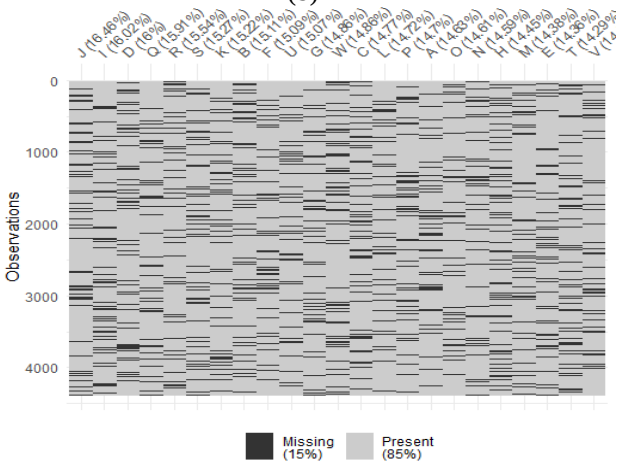
Dataset used: Spellman				
Percentage of missing data	5	10	15	20
Number of missing values	5038	10076	15114	20152
Number of complete values	95725	90687	85649	80611
Performance analysis of various missing data Imputation Technique	Percentage of missing data			
	5	10	15	20
(RMSE value)				
mean	0.1174	0.1172	0.1168	0.1169
median	0.1174	0.1173	0.1169	0.117
mode	0.1192	0.1199	0.1196	0.1237
decision	0.0042	0.0035	0.0029	0.0044
KNN	0.0071	0.0056	0.0047	0.0042
PMM	0.0023	0.0015	0.0012	0.0015
BLR	0.002	0.0017	0.0013	0.0015
R.FOREST	0.0039	0.0028	0.0024	0.0021



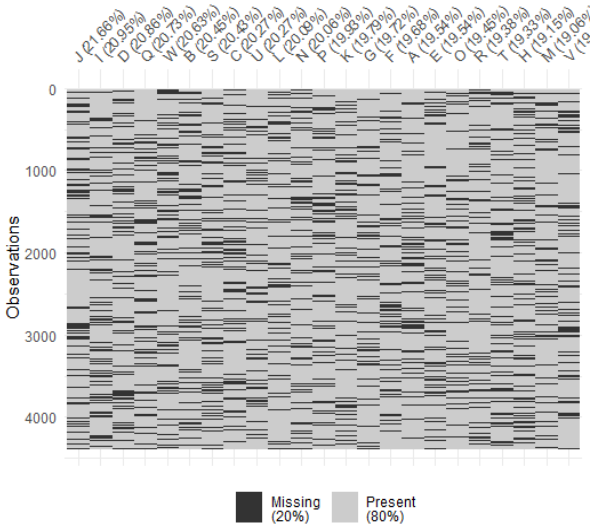
(a)



(b)



(c)



(d)

Figure. 2 Percentage of missing values pattern in the Spellman data set a)5% b)10% c)15% d)20%.

V. CONCLUSION

Missing data values is an unwanted issue in recent days. From the history of data science, many authors have proposed different techniques to overcome this missing data values issue. This missing data can be estimated either as a single value or multiple values. In our study, we have made a comparative study of mean, mode, median, KNN, PMM, BLR and Random-Forest. The mean, mode, and median are the traditional way of imputation algorithms which can leads to bias, whereas modern approach like KNN which considers the nearest observations to impute in the place of missing data, but it has some drawbacks like search space and how many nearest neighbors are required for KNN is not known prior. PMM is a multiple imputation algorithm where it first identifies the linear regression coefficients then from the coefficients it estimates the missing values. BLR this techniques identifies the missing values from the distribution, it takes help from regression coefficients. Random-Forest in this given data set is formed as many as trees and each tree and its performance is evaluated, and at the end, it selects the most appropriate tree for the imputation.

REFERENCES

1. Alexandre G de Brevern, Serge Hazout and Alain Malpertuy " Influence of microarrays experiments missing values on the stability of gene groups by hierarchical clustering" BMC Bioinformatics 2004, 5:114
2. Xian Wang, Ao Li, Zhaohui Jiang and Huanqing Feng " Missing value estimation for DNA microarray gene expression data by Support Vector Regression imputation and orthogonal coding scheme" BMC Bioinformatics 2006, 7:32 doi:10.1186/1471-2105-7-32
3. Jianjun Hu, Haifeng Li, Michael S Waterman and Xianghong Jasmine Zhou " missing value estimation for microarray data" BMC Bioinformatics 2006, 7:449
4. Dankyu Yoon, Eun-Kyung Lee and Taesung Park " Robust imputation method for missing values in microarray data" BMC Bioinformatics 2007, 8(Suppl 2):S6
5. Huihui Li1, Changbo Zhao, Fengfeng Shao, Guo-Zheng Li, Xiao Wang " A hybrid imputation approach for microarray missing value estimation" From IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2014) Belfast, UK. 2-5 November 2014
6. Hsiuying Wang, Chia-Chun Chiu, Yi-Ching Wu, Wei-Sheng Wu " Shrinkage regression-based methods for microarray missing value imputation" From 24th International Conference on Genome Informatics (GIW 2013) Singapore, Singapore. 16-18 December 2013
7. Fanchi Meng, Cheng Cai, and Hong Yan " A Bicluster-Based Bayesian Principal Component Analysis Method for Microarray Missing Value Estimation" ieee journal of biomedical and health informatics, vol. 18, no. 3, may 2014
8. Ming Ouyang, William J. Welsh and Panos Georgopoulos " Gaussian mixture clustering and imputation of microarray data" Bioinformatics Vol. 20 no. 6 2004, pages 917–923
9. Rubin, D. B. (1976). Inference and missing data. Biometrika 63(3): 581-592.
10. Rubin, D. B. (1987). Multiple Imputations for Nonresponse in Surveys. New York, J. Wiley & Sons.
11. Enders, C. K. (2010). Applied Missing Data Analysis. New York, NY, The Guilford Press.
12. Baraldi, A. N. and C. K. Enders (2010). An introduction to modern missing data analyses. Journal of School Psychology 48(1): 5-37.
13. Preeti Patida and Anshu Tiwar " Handling Missing Value in Decision Tree Algorithm" International Journal of Computer Applications (0975 –8887)Volume 70–No.13, May 2013
14. Lorenzo Beretta and Alessandro Santaniello "Nearest neighbor imputation algorithms: a critical evaluation" BMC Medical Informatics and Decision Making ,20 16 (Suppl 3) :74
15. Morris, Tim P., Ian R. White and Patrick Royston (2014) "Tuning multiple imputation by predictive mean matching and local residual draws." BMC Medical Research Methodology 14: 75-87.
16. Rubin, Donald B. (1987) Multiple Imputation for Non response in Surveys. Wiley.
17. L. Breiman. Random forests. Machine learning , 45(1):532, 2001. ISSN 0885-6125.
18. Paul T. Spellman,Gavin Sherlock, Michael Q. Zhang, Vishwanath R. Iyer, Kirk Anders, Michael B. Eisen, Patrick O. Brown, David Botstein, and Bruce Futcher "Comprehensive Identification of Cell Cycle–regulated Genes of the Yeast Saccharomyces cerevisiae by Microarray Hybridization" Molecular Biology of the CellVol. 9, No. 12.
19. Dinesh Singh, et al. "Gene Expression Correlates of Clinical Prostate Cancer Behavior". Cancer Cell, 1:203-209, March, 2002
20. Alan Wee-Chung Liew Ngai-Fong Law Hong Yan " Missing value imputation for gene expression data: computational techniques to recover missing data from available information" Brief Bioinform. 2011 Sep;12(5):498-513

AUTHORS PROFILE

K.Ishthaq Ahamed received M.Tech from Indian School of Mines, Dhanbad and presently working as Associate Professor in Computer Science and Engineering Department in G Pulla Reddy Engineering College, Kurnool.

Dr. Shaheda Aktha received Bachelor of Computer Science, Master of Computer Science from Acharya Nagaurjuna Universtiy, M.S from B.I.T.S Pilani and Ph.D from Acharya Nagarjua University. Presently working as Registrar F.A.C in Dr. Abdul Haq Urdu University, Kurnool.