# Feature Extraction In Gene Expression Dataset Using Multilayer Perceptron

**Nageswara Rao Eluri[1], Gangadhara Rao Kancharla[2], Suresh Dara[3]**

[1]Department of CSE, Acharya Nagarjuna University, Guntur, AP, India-522510.
eluri76@gmail.com,kancherla123@gmail.com
[2]Department of CSE, Acharya Nagarjuna University, Guntur, AP, India-522510.
[3]Department of CSE, B V Raju Institute of Technology, Narsapur, Telangana, India-502313. darasuresh@live.in

**Article History**: Received: 11 January 2021; Accepted: 27 February 2021; Published online: 5 April 2021

**Abstract**:Numerous amount of gene expression datasets that are publicly available have accumulated since decades. It is hence essential to recognize and extract the instances in terms of quantitative and qualitative means.In this study, Keras is utilized to model the multilayer perceptron (MLP) to extract the features from the given input gene expression dataset. The MLP extracts the features from the test datasets after its initial training with the top extracted features from the training classifiers. Finally with the top extracted features, the MLP is fine tuned to extract optimal features from the gene expression datasets namely Gene Expression database of Normal and Tumor tissues 2 (GENT2). The experimental results shows that the proposed model achieves better feature selection than other methods in terms of accuracy, f-measure, precision and recall.

**Keywords**: Gene Expression Dataset, Deep Learning, Multilayer Perceptron, Keras

## 1. Introduction

The importance of microarrays in biomedical and biological science has recently come to be increased [1,2]. The advent of microarray technologies has contributed to advances in this technology. This modifies research into multiple gene studies under different circumstances and allows detailed data to be analysed.

The key approach taken in the analysis of the resulting knowledge is the clustering procedure, among the main strategies considered. Certainly, genes had initially reacted transcript in some experimental circumstances. For their performance in the discovery of the genes in certain cases some strategies are perceived. In all cases, the subset of genes which are associated under few sub-sets conditions is difficult to discover. Moreover, no further clusters of the given genes [3] are allocated. In addition, certain sub-sets of genes have comparative behaviors, which give individual behaviors under other conditions [4].

The researchers also started the clustering process [5] with a view to reducing the disadvantages associated with the processes of gene expression data collection [6]. The clustering method includes deciding the genes with the same activity have a category or grouping that are accessible under such circumstances. That's why the NP-Hard is considered. A variety of methods are available to tackle the issue and explore the search field using machine learning models [7] [8].

In this paper, Keras is utilized to model the multilayer perceptron (MLP) to extract the features from the given input gene expression dataset. The MLP extracts the features from the test datasets after its initial training with the top extracted features from the training classifiers. MLP is fine-tuned with top extracted features to extract optimal features GENT2 datasets.

## 2. Background Study

The survey of biclustering solutions with statistical signification, despite increasing contribution to biclustering approaches, is poorly examined [9]. We discuss that this is the case and therefore consider the main limitations of the current methods.

Hybrid and deep approaches have been applied by Aziguli et al. [15] to reduce noise and to increase extraction performance. Similarly Jiang et al. [16] proposed that text be classified in a sparse matrix for the text clustering challenge, by means of a text clustering model for extraction, in order to solve the computation problem.

The hybrid deep-belief algorithm for feeling clustering was suggested by Edinburgh et al. [22]. First they extracted the characteristics from the previously hidden Boltzmann layers in their two fold networks using the Convolutional Restricted Boltzmann Machines.

In order to learn emotional properties from speech patterns, Huang et al. [23] used deep-belief network. A grouping of non-linear SVMs has been used to create a hybrid emotional detection procedure using extracted characteristics.

Kahuet al. [24] shows that a further change is possible in the decay of RELU units rather than max out units. Liu et al. [25] is careful about the significance of corpus words in clustering of neural networks.

Stochastic approaches to biclustering are primarily based on multivariate evidence [10]. Educated results, however, are not used to assess the importance of biclusters. Instead, bicluster derivation is rendered by qualified data where clear convergence conditions are met.

Methodologies for clustering [11] define cluster target homogeneity metrics for guidance of exploration, sort the biclusters detected and then channel the biclusters. The features of the sample data do not guarantee the identification of clusters. This little clusters are found to be highly homogeneous.

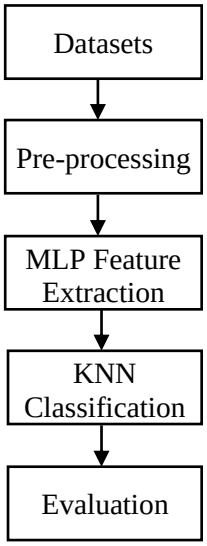
This unnecessary effect, including the size of the bicluster, is balanced by merit functions and benefits large quantities of biclusters [12]. The value of the bicluster is not properly assured and this promotes high-risk detection of poor homogeneity of false positive genes.

In order to guarantee a greater residual fault than homogeneity with a particular statistic strength and meaning, the statistical test is also applied [13]. The compactness of the bicluster is ensured when certain columns or rows are included or excluded in the gene matrix, which tends to improve the homogeneity of gene expression. This technique is, however, influenced by an issue of homogeneity that does not exclude biclusters [14].

These methods handle continuous coherence in a certain manner, even with the significant work on existing literatures, and therefore are not suitable for assessing biclustering solutions generated by optional algorithms.

## 3. Research work

A multi-objective optimisation and a thorough evaluation for excluding the irrelevant matrix or those with a smaller correlating is proposed in this paper with the Keras tool. Initially, it incorporates a pre-processing model to remove unwanted elements and then text feature extraction using MLP. Finally, KNN is used for classifying the instances from the gene expression datasets




## 4. Methodologies

This section shows the details of pre-processing, feature extraction using MLP and modelling of MLP for gene expression datasets.

### 1.1. Data and Pre-processing

To evaluate and demonstrate the application of our model, we used a GENT2 dataset.

### 1.2. Multilayer Perceptron

A Multilayer Perceptron (MLP) is a network that tries to map input onto output. An MLP has multiple layers of nodes where each layer is fully connected with the next one. Each node of the hidden layers is operated with a

nonlinear activation function. A backpropagation algorithm is used to train the network. Now let's explain activation functions for training and learning through backpropagation.

MLP is a feedfoward neural network that attempts to map the input into the output. An MLP has several neural network layers where each layer is connected with next layer. A nonlinear activation function is used at the each node of hidden layers. The network is trained using a backpropagation algorithm and activation features for backpropagation preparation and learning is given below.

The linear regression model of the final layer of our model can be represented as Equation 1:

$$f(x) = w^T x + b \quad (1)$$

where

$x$ – input

$w$ – weighted matrix

$b$ – bias that gets trained to reduce the activation function.

***Activation Function:***

The study uses two different activation function for the purpose of training. The first activation function is a hyperbolic tangent, where its evaluation ranges between -1 and 1, and it is described as below:

$$y(v_i) = \tanh(v_i) \quad (2)$$

Secondly, the study uses a logistic function with its evaluation ranges between 0 and 1 and it is described as below:

$$y(v_i) = \left(1 + e^{-v_i}\right)^{-1} \quad (3)$$

where

$y_i$- output of $i^{th}$ neuron and

$v_i$- weighted sum of input.

***Backpropagation Learning:***

After analyzing the data for each neuron, an MLP network can be trained by adjusting link weights. The study conducts supervised learning regardless of the amount of error in the output relative to the predicted outcomes. In the study the errors $e$ are quantified in a node of $n^{th}$ row of the training data as below:

$$error_j(n) = d_j(n) - y_j(n) \quad (4)$$

where

$d$- Expected value and

$y$ - Target value.

The model makes suitable corrections in order to reduce the probability of errors on the MLP output as below:

$$e(n) = 0.5\sum_i error_j^2(n) \quad (5)$$

The change in weight after the application of gradient descent is represented as below:

$$\Delta w_{ji}(n) = -\eta \frac{\partial e(n)}{\partial v_j(n)} y_i(n) \quad (6)$$

where

$y$- Previous layer output

$\eta$ - Learning rate or momentum.

With several induced local fields, the study tends to define a derivative for the overall output node and it is defined as below:

$$-\frac{\partial e(n)}{\partial v_j(n)} = error_j(n)\phi'\left(v_j(n)\right) \quad (7)$$

where

$\phi'$ - activation function derivative with constant rate.

If there exist a change in weights in the hidden layers, the analysis tends to become difficult and hence it is necessary to provide the following expression of a relevant derivative to ensure the analysis becomes easier.

$$-\frac{\partial e(n)}{\partial v_j(n)} = \phi'\left(v_j(n)\right)\sum_k -\frac{\partial e(n)}{\partial v_k(n)} w_{kj}(n) \quad (8)$$

This relative derivative depends on the change in node weights, which is represented in the output layer. Therefore, in order to change the hidden layer weights, we must first change the output layer weights according to the derivative of the activation function. This analysis thus represents a backpropagation of the activation function.

This relative derivative is dependent on the shift in the node weight in the output layer. Therefore, the ML tends to adjust the output layer weights according to the activation function derivative in order to change hidden layer weights. This study thus represents a backpropogation in its activation mechanism.

### 1.3. Modelling MLP for Gene Expression Dataset

The fitness function is thus formulated as,

$$f_1(x) = \frac{GC}{size(x)} \tag{15}$$

$$f_2(x) = \frac{H(x)}{\delta} \tag{16}$$

where

*G*- total sub-matrix rows and

C - sub-matrix columns of a *x*matrix.

size(*x*) - capacity of matrix *x*, which is the product of rows and columns.

*H*(*x*) – threshold value or mean secured error sub-matrix *s*that belongs to the matrix *x*.

A point eliminating method reduces the mean squared residue in the size(*x*) with maximum MSR. The capacity of *s*sub-matrix is hence increased prominently to ensure that the value of MSR liesbehind the user defined threshold.

The objective function for finding the features is hence formulated as below:

$$F(G,C) = \begin{cases} f(G,C)^{-1} & if\ H(x) \le \delta \\ H(x) & otherwise \end{cases} \tag{17}$$

## 5. Results and Discussion

This section describes the evaluation of the model for evaluating the proposed MLP for feature extraction over GENT2 datasets. The GENT2 is updated at regular intervals with the gene expression patterns across that consisting of normal and tumor tissues from public gene expression data sets.

The performance is estimated in terms of accuracy, sensitivity, specificity, f-measure, percentage error and geometric mean represented between Fig.2 – Fig.7.

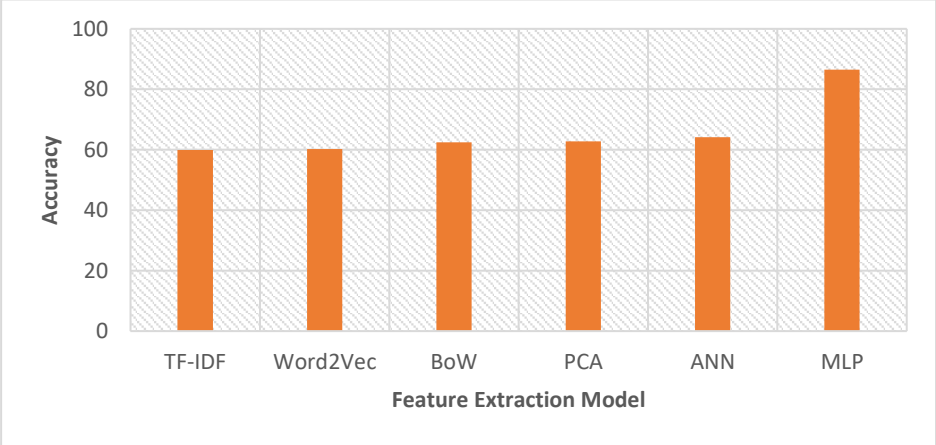



Fig.2. Accuracy

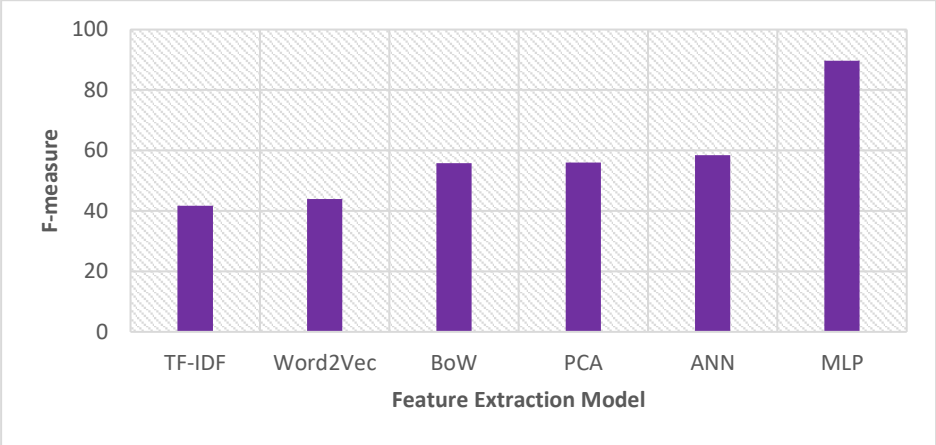



Fig.3. F-measure

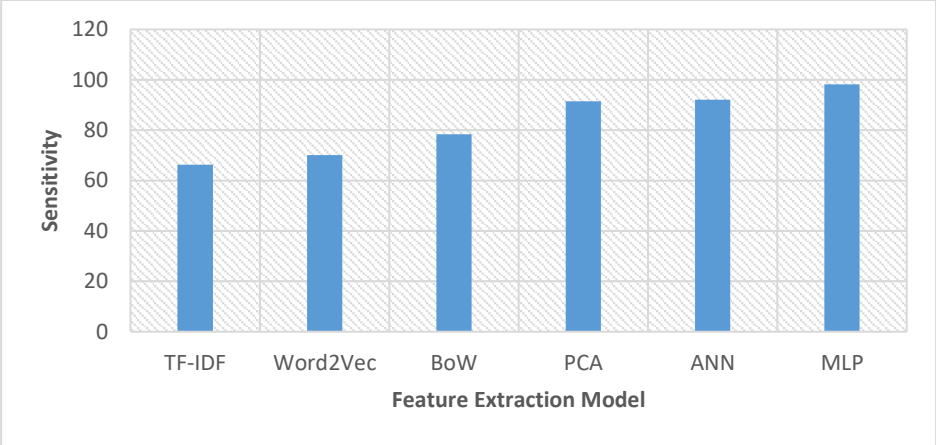



Fig.4. Sensitivity

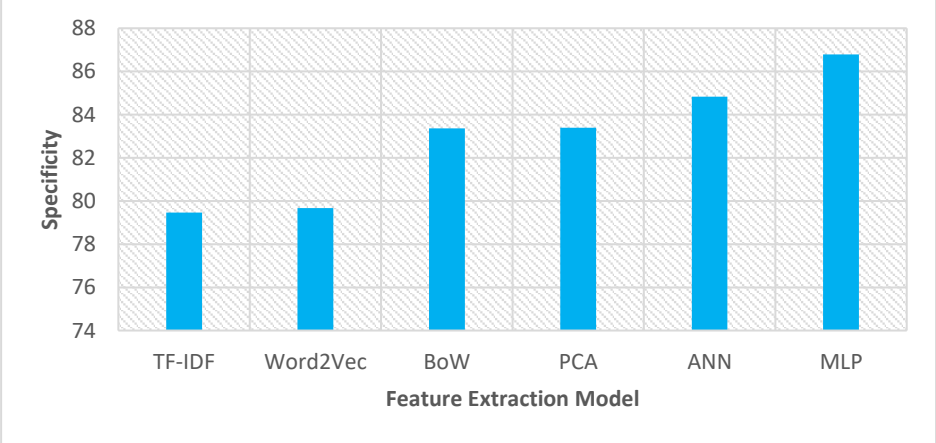



Fig.5. Specificity

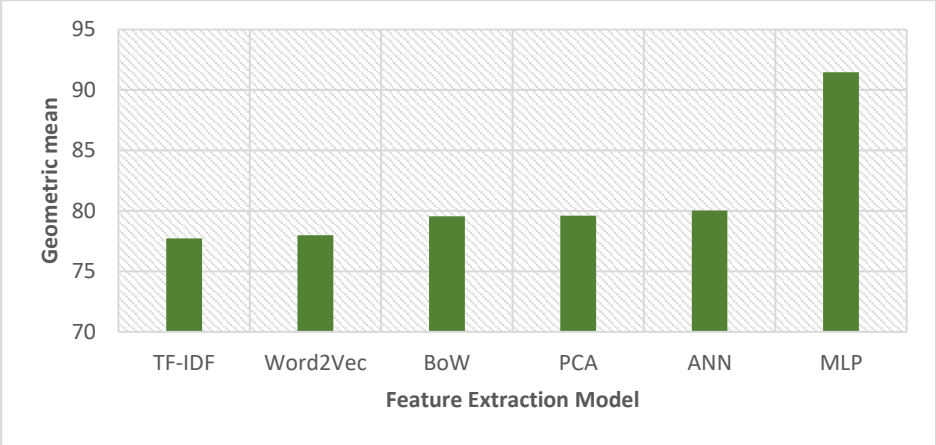



Fig.6. Geometric mean

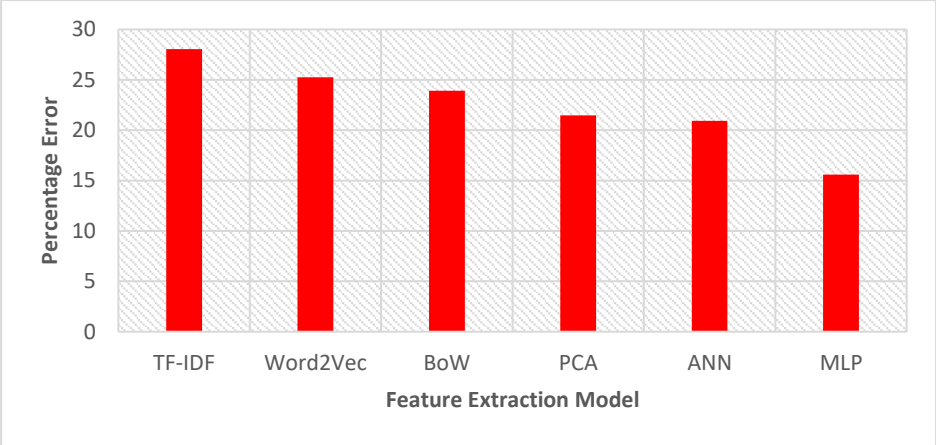



Fig.7. Percentage error

From the results of Fig.2 – Fig.7, the results shows of classification results from the feature extracted using MLP. The comparison of results is made between various feature extraction models like TF-IDF, word2vec, bag of weights, principle component analysis, artificial neural network and multi layered perceptron. The classification for all these feature extraction models are carried out using KNN classifier. The results of simulation shows that the proposed MLP model attains increased classification accuracy, f-measure, sensitivity, specificity, geometric mean and reduced percentage error than existing text feature extraction methods.

## 6. Conclusion

In this paper, Keras modelling on MLP extract essential features from the input gene expression dataset. The MLP extracts the features from the test datasets after its initial training with the top extracted features from the training classifiers. Finally with the top extracted features, the MLP is fine tuned to extract optimal features from the gene expression datasets namely GENT2. The results shows that the MLP extracts well the feature than other methods in terms of accuracy, f-measure, precision and recall.

**References**

1. Tanay, A., Sharan, R., & Shamir, R. (2002). Discovering statistically significant biclusters in gene expression data. Bioinformatics, 18(suppl_1), S136-S144.
2. Sancetta, A. (2016). Greedy algorithms for prediction. Bernoulli, 22(2), 1227-1277.
3. de França, F. O., Bezerra, G., & Von Zuben, F. J. (2006, July). New perspectives for the biclustering problem. In Evolutionary Computation, 2006. CEC 2006. IEEE Congress on (pp. 753-760). IEEE.
4. Yip, K. (2003). DB seminar series: Biclustering methods for microarray data analysis, 46–47, 2003.
5. Hartigan, J. A. (1972). Direct clustering of a data matrix. Journal of the american statistical association, 67(337), 123-129.
6. Cheng, Y., & Church, G. M. (2000, August). Biclustering of expression data. In Ismb (Vol. 8, No. 2000, pp. 93-103).
7. Ayadi, W., Elloumi, M., & Hao, J. K. (2009). A biclustering algorithm based on a Bicluster Enumeration Tree: application to DNA microarray data. BioData mining, 2(1), 9.

8. Ayadi, W., Elloumi, M., & Hao, J. K. (2012). BicFinder: a biclustering algorithm for microarray data analysis. Knowledge and Information Systems, 30(2), 341-358.
9. Henriques, R. (2016). Learning from high-dimensional data using local descriptive models (Doctoral dissertation, PhD thesis, Instituto Superior Tecnico, Lisboa: Universidade de Lisboa).
10. Hochreiter, S., Bodenhofer, U., Heusel, M., Mayr, A., Mitterecker, A., Kasim, A., ... &Bijnens, L. (2010). FABIA: factor analysis for bicluster acquisition. Bioinformatics, 26(12), 1520-1527.
11. Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB), 1(1), 24-45.
12. Mitra, S., & Banka, H. (2006). Multi-objective evolutionary biclustering of gene expression data. Pattern Recognition, 39(12), 2464-2477.
13. Wang, H., Wang, W., Yang, J., & Yu, P. S. (2002, June). Clustering by pattern similarity in large data sets. In Proceedings of the 2002 ACM SIGMOD international conference on Management of data (pp. 394-405). ACM.
14. Chandragandhi, S., Raja, R. A., Dhiman, G., & Kaur, A. (2021). Analysis of protein-ligand interactions of SARS-Cov-2 against selective drug using deep neural networks. Big Data Mining and Analytics, 4(2), 76-83.
15. Aziguli, W., Zhang, Y., Xie, Y., Zhang, D., Luo, X., Li, C., & Zhang, Y. (2017). A robust text classifier based on denoising deep neural network in the analysis of big data. *Scientific Programming*, *2017*.
16. Jiang, M., Liang, Y., Feng, X., Fan, X., Pei, Z., Xue, Y., & Guan, R. (2018). Text classification based on deep belief network and softmax regression. *Neural Computing and Applications*, *29*(1), 61-70.
17. Kousik, N., Natarajan, Y., Raja, R. A., Kallam, S., Patan, R., &Gandomi, A. H. (2021). Improved salient object detection using hybrid Convolution Recurrent Neural Network. Expert Systems with Applications, 166, 114064.
18. Karthikeyan, T., &Praghash, K. (2020). An Improved Task Allocation Scheme in Serverless Computing Using Gray Wolf Optimization (GWO) Based Reinforcement Learning (RIL) Approach. Wireless Personal Communications, 1-19.
19. Daniel, A., Kannan, B. B., Yuvaraj, N., &Kousik, N. V. (2021). Predicting Energy Demands Constructed on Ensemble of Classifiers. In Intelligent Computing and Applications (pp. 575-583). Springer, Singapore.
20. Srihari, K., Dhiman, G., Somasundaram, K., Sharma, A., Rajeskannan, S., ... &Masud, M. (2021). Nature-Inspired-Based Approach for Automated Cyberbullying Classification on Multimedia Social Networking. Mathematical Problems in Engineering, 2021.
21. Sangeetha, S. B., Blessing, N. W., & Sneha, J. A. (2020). Improving the training pattern in back-propagation neural networks using holt-winters' seasonal method and gradient boosting model. In Applications of Machine Learning (pp. 189-198). Springer, Singapore.
22. Zhou, S., Chen, Q., & Wang, X. (2014). Active semi-supervised learning method with hybrid deep belief networks. *PloS one*, *9*(9), e107122.
23. Huang, C., Gong, W., Fu, W., & Feng, D. (2014). A research of speech emotion recognition based on deep belief network and SVM. *Mathematical Problems in Engineering*, *2014*.
24. Kahou, S. E., Pal, C., Bouthillier, X., Froumenty, P., Gülçehre, Ç.,Memisevic, R., ... &Mirza, M. (2013, December). Combining modality specific deep neural networks for emotion recognition in video. In *Proceedings of the 15th ACM on International conference on multimodal interaction* (pp. 543-550).
25. Liu, M., Haffari, G., Buntine, W., &Ananda-Rajah, M. (2017, December). Leveraging linguistic resources for improving neural text classification. In *Proceedings of the Australasian Language Technology Association Workshop 2017* (pp. 34-42).
26. Gangadhar Rao Kancharla Suresh dara, Priyanka Tumma, Nageswara Rao Eluri.(2018)Feature Extraction in Medical Images by using Deep Learning Approach. International Journal of Pure and Applied Mathematics.
27. Gangadhara Rao Kancharla, Nageswara Rao Eluri, Suresh Dara, Nishath Ansari (Mar-2019). An efficient algorithm for feature selection problem in gene expression data: A spider monkey optimization approach. Proceedings of 2nd International Conference on Advanced Computing and Software Engineering (ICACSE)
28. Suresh Dara, Mamidi Jagadeeshwara Reddy, Nageswara Rao Eluri (Mar-2018). Evolutionary Computation based Feature Selection: A Survey, 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA).
29. Dr.A.Senthil Kumar, Dr.G.Suresh, Dr.S.Lekashri, Mr.L.Ganesh Babu, Dr. R.Manikandan. (2021). Smart Agriculture System With E – Carbage Using Iot. International Journal of Modern Agriculture, 10(01), 928 - 931. Retrieved from http://www.modern-journals.com/index.php/ijma/article/view/690
30. Dr.G.Suresh, Dr.A.Senthil Kumar, Dr.S.Lekashri, Dr.R.Manikandan. (2021). Efficient Crop Yield Recommendation System Using Machine Learning For Digital Farming. International Journal of Modern

Agriculture, 10(01), 906 - 914. Retrieved from http://www.modern-journals.com/index.php/ijma/article/view/688

31. Dr. R. Manikandan, Dr Senthilkumar A. Dr Lekashri S. Abhay Chaturvedi. “Data Traffic Trust Model for Clustered Wireless Sensor Network.” INFORMATION TECHNOLOGY IN INDUSTRY 9.1 (2021): 1225–1229. Print.