Webology (ISSN: 1735-188X)
Volume 18, Number 5, 2021

Enhancing Breast Cancer Diagnosis: A Neural Network-Based Clustering Approach For Segmentation

Dr. T. S. Poornappriya[1], Dr. R. Gopinath[2]

[1]Data Scientist, Tech Mahindra Bengaluru, Karnataka, India.

[2]D.Litt. (Business Administration) - Researcher, Madurai Kamaraj University, Madurai, Tamil Nadu, India.

Abstract

The second most common type of cancer that kills women is breast cancer. The greatest tool for early breast cancer screening is mammography, which can spot tiny tumours up to two years before they become noticeable by physical examination. To spot early indications of malignant growth, X-ray images of the breast must be carefully studied. During the process of image processing and interpretation, radiographic pictures are typically segmented or partitioned into parts of similar texture. Segmentation is incredibly difficult in mammographic pictures because of the relative lack of structure definition and the implied transition from one texture to another. Since it is unknown how many different regions there are in the image, the task of examining various texture sections might be seen as an exploratory report. In this study, a segmentation technique based on SOM is presented.

Keywords: Breast Cancer, Mammography, Self-Organizing Map, Euclidean Distance, Validity Measure, Double Bouldin Index.

1. Introduction

Breast cancer is the second most frequent type of cancer in women, yet it continues to be the top cause of cancer mortality in women between the ages of 40 and 55, according to the USA Cancer Society. Invading breast cancer will be discovered in about 180,200 women in the United States this year. During the same year, around 44,190 women will succumb to this fatal illness [1]. Although the percentage of new cases of breast cancer increased by an average of 4% between 1982 and 2017, the percentage rate has recently tapered off to little over 1%. The increased use of mammography to identify the disease in its early stages has been credited with a large portion of this welcome decline in new breast cancer diagnoses. Although mammography technology has made great advancements in prediction techniques, much work has to be done to increase overall detection accuracy [2].

Segmentation is the division of an image into several pieces [3]. A region's pixels all have the same qualities, such as colour, intensity, or texture. Parallel computing models called artificial neural networks [4] are made up of tightly coupled adaptive processing units. These networks' ability to learn by doing is a key characteristic.

Artificial neural networks are better suited for situations where training data is easily accessible but where one has a limited or imperfect grasp of the problem to be solved due to their adaptive nature.

2. Self-Organizing Map (SOM)

The high-dimensional data is projected onto the two-dimensional map using a self-organizing map [5]. The dimensional reduction might make it easier for us to see the key relationship between the data. SOM is famous for having the topology structure property that is seen in the brain but not in any other artificial neural network. Since it keeps the neighbourhood relation of the input pattern, it is referred to as topology-preserving. Units that are physically close to one another will react to input vector classes that are also close to one another.

Two layers—an input layer and an output layer—make up the fundamental SOM model. Neurons in the SOM network resemble brain neurons in structure and function. All of the units are fully connected to each input. The number of clusters in the image that needs to be segmented determines the number of neurons in the output layer, hence the number of clusters is equal to the number of output neurons. One of the key characteristics utilised for image segmentation is colour. SOM is used to translate color-space patterns from three dimensions to two dimensions. SOM gains knowledge from competition. Only one neuron in the network will react to each input vector. The mechanism behind this is called competition. The weights of that neuron and the neurons around are updated after a neuron is selected as the winner. The SOM neighbourhood plan may be circular, hexagonal, or rectilinear. The multicomponent values are provided as training input. The neighbourhood size is initially set to the maximum of either the network's height or breadth divided by two, and the learning rate is initially set to 0.1. The neurons' weight vectors are initialised at random. The input vectors that need to be clustered are supplied to the network in a random sequence for each iteration. The winner or best matching unit is determined to be the neurons whose weight vector best matches the input vector (BMU). Using the Euclidean distance approach, the winner is decided as follows:

$$\left\|x - W_l^{[k]}\right\| = \min_i \left\|x - W_i^{[k]}\right\| \qquad (1)$$

W is the weight of the victorious unit I at each iteration k, where x is the input vector. The weights of the winning neuron and the neurons in its immediate vicinity are changed to

bring them closer to the input vector being fed into the network. The updated weights are listed below.

$$W_i^{k+1} = W_i^k + H_{li}^k (x - W_i^k) \qquad (2)$$

where H represents the smoothing kernel over the victorious neuron. With regard to the Gaussian function, the kernel can be expressed as

$$H_{li}^k = \alpha^k exp\left(-\frac{d^2(l,i)}{2(\sigma^k)^2}\right) \qquad (3)$$

where d is the neighbourhood distance, I is the winning neuron, d is the distance between them, and k is the learning rate at iteration k. After each cycle, the neighbourhood size and learning rate are updated. The neighbourhood and learning rate both decline as the number of iterations rises. As shown below, the learning rate is exponentially lowered:

$$\alpha^k = \alpha^0 \exp\left(-\frac{k}{T}\right) \qquad (4)$$

where T is the total number of iterations, which is set to 1000, and σ^0 is the initial learning rate. The neighborhood's decreasing function is provided as follows:

$$\sigma^k = \sigma^0 \left(1 - \frac{k}{T}\right) \qquad (5)$$

where k is the number of iterations and σ^k is the size of the neighbourhood at the beginning. The neighbourhood σ^0 gets smaller until it only includes one house.

The input is translated from a high colour space to a two-dimensional map after the SOM converges. The final outcome of SOM is dependent on the starting weight values, training data, and map properties like certain network nodes, learning rate, and neighbourhood. SOM has a problem with over-segmentation. In order to determine the ideal number of clusters, an optimization technique like the genetic algorithm is used. An optimization strategy is used to locate the cluster centres using the data set discovered by SOM as an input.

3. Proposed Framework for the Segmentation of Breast Cancer Images

The proposed framework for classifying breast cancer cells utilising the pre-processing and segmentation steps is shown in figure 1 below. The pre-processing step in this work uses the median filtering approach to eliminate image noise, while the segmentation step uses SOM.

3.1 Pre-Processing with Median Filter

An undesired signal in the picture is called noise. There are three sorts of noise in the image. Gaussian noise, impulse noise, and salt-and-pepper noise. An advantage of employing a median filter is that it has a robust average, which means that an unrepresentative pixel in the neighbourhood has no effect on the median value, and it also has the ability to preserve sharp edges. A median filter works by choosing the median intensity in the window. The window moves pixel by pixel across the entire image as the median filter moves from pixel to pixel through it, replacing each one with the median value of its neighbours.

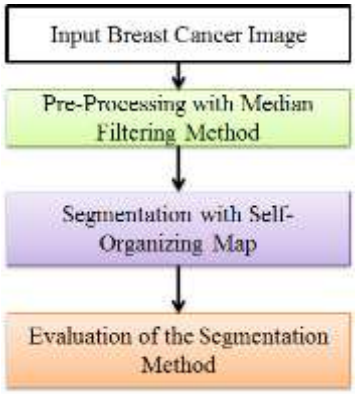


Figure 1: Proposed Framework for the SOM based Segmentation of Breast Cancer Image

3.2 Segmentation using SOM

In this work, the SOM approach uses the grey scale values of each pixel as an input during the clustering phase. In this study, the neighbourhood topology used in the SOM approach is a linear array, sometimes referred to as a one-dimensional (1-D) topology. The SOM algorithm's calculation is divided into two stages: the learning stage and the recognition stage. This study employs Normalized Euclidean Distance rather than Euclidean Distance to calculate distance.

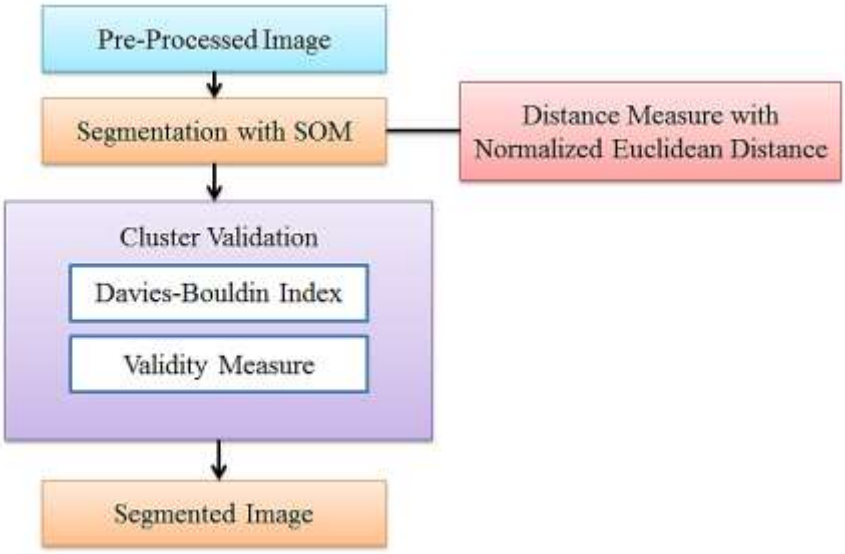


Figure 2: Segmentation Process of Self-Organizing Map

3.2.1 Normalized Euclidean Distance

The Euclidean Distance is adjusted in the computation of the Normalized Euclidean Distance [8]. The following equation illustrates the normalised Euclidean distance of two vectors, between vector u and vector v.

$$d_{(uv)} = \sqrt{\sum_{k=1}^{n} (\overline{u} - \overline{v})^2}$$

Where

$$\overline{u_i} = \frac{u_i}{|u|}, \quad \overline{v_i} = \frac{v_i}{\|v\|}$$

The normalised value of the vector v is ||v||. The following equation represents the normalised value.

$$\|v\| = \left[\sum_{i=1}^{n} v_i^2\right]^{\frac{1}{2}}$$

3.2.2 Cluster Validation:

A. Davies-Bouldin Index (DBI)

David L. Davies and Donald W. Davies preceded DBI in 1979. The outcomes of clustering are evaluated using DBI [9]. The DBI approach compares the overall within-cluster scatter

(a cluster's spread) to the between-cluster separation (distance between clusters). The spread of cluster value is calculated using the equation below.

$$S_i = \frac{1}{T_i} \sum_{x \in C_i} \|x - z_i\|$$

where T_i is the total number of members in cluster (C_i), and z_i is the centre of that particular cluster. The Euclidean distance between the centres of the i^{th} and j^{th} clusters is used to compute the distance between clusters. Its distance is calculated using the calculation below.

$$d_{ij} = \|z_i - z_j\|$$

The ratio value between the i^{th} and j^{th} clusters is known as R_{ij} and is determined using the formula below.

$$R_{ij} = \left\{ \frac{S_i + S_j}{d_{ij}} \right\}$$

Finding the ratio's highest value (D_i) is used to determine the value of DBI. Calculating the value of D_i is done using the equation below.

$$D_i = \max_{j; j \neq i} R_{ij}$$

Then, Equation is used to calculate the DBI value.

$$DBI = \frac{1}{K} \sum_{i=1}^{K} D_i$$

k is the number of clusters, in this case.

The DBI with the lowest value demonstrates the best clustering outcomes and produces a well-separated cluster.

B. Validity Measure (VM)

One index used to evaluate the reliability of clustering findings is VM [10]. The application of image segmentation based on clustering frequently uses virtual machines (VM). The equation shown below is used to compute VM.

$$VM = y\left(\frac{intra}{inter} \right)$$

where intra is the distance within a cluster, inter is the distance between clusters, and y is a function of the number of clusters that are formed. To determine the value of intra-cluster distance, apply the equation below.

$$intra = \frac{1}{N}\sum_{i=1}^{k}\sum_{x:C_i}\|x - z_i\|^2$$

where k is the number of clusters, z_i is the centre of cluster C_i, and N is the total number of pixels in the image.

Additionally, the minimum inter-cluster distance value [10] is used to calculate virtual machine. To determine the inter-cluster distance, apply the equation below.

$$inter - \min(\|z_i - z_j\|)$$

where I = 1, 2,..., k and j = i+1,..., k. y is multiplied by the ratio of the intra-cluster and inter-cluster distances. To compute y, use the equation below.

$$y = c.N(2,1) + 1$$

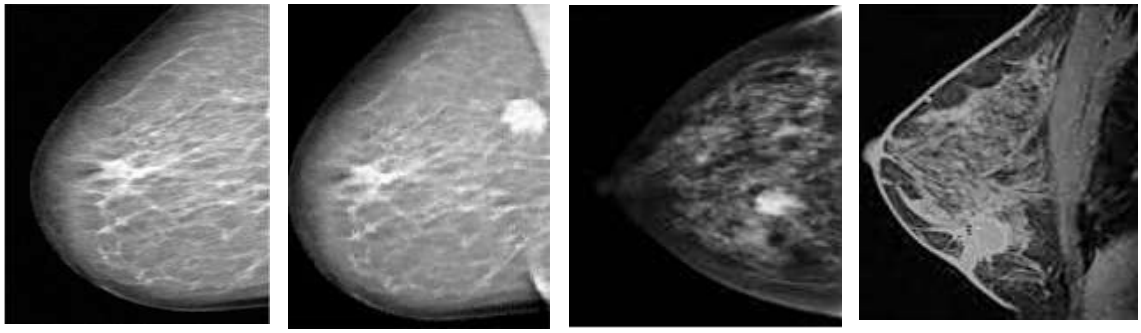
where N (2,1) is a Gaussian function for the number of clusters and c is a constant between 15 and 25, (k). Equation displays the Gaussian function.

$$N(\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left|\frac{(k-\mu)^2}{2\sigma^2}\right|}$$

To achieve the best outcome and a well-separated cluster, VM should be kept to a minimum.

4. Result and Discussion

Following figure 3 presents the breast cancer images considered for the segmentation process whereas a) Image1.jpg, b) Image2.jpg, c) Image3.jpg, d) Image4.jpg and table 1 gives the initialization of parameters on SOM method and spatial operations



a) Image1.jpg b)Image2.jpg c)Image3.jpg d)Image4.jpg

Figure 3: **Breast Cancer Image for Segmentation process a) Image1.jpg b) Image2.jpg c) Image3.jpg d) Image4.jpg**

Table 1: **Initialization of parameters on SOM Method**

Sl.No	Parameter	Value
1	A	0.2
2	Epoch (T)	200
3	The radius of the noise filter	3
4	Threshold Region (ThA)	0.003 * total pixels

Tests are performed on each image by forming 2 until 10 clusters. From each cluster which is formed, then calculate the value of Validity Measure (VM) and Davies-Bouldin Index (DBI). Each value of VM and DBI with minimum value showed the most optimal number of cluster. Results of validity measurement for each test images are shown in the following table2a – table 2d.

Table 2a: Validity Measure and Davis-Bouldin Index for the given breast cancer Image1.jpg based on the number of clusters

Number of Clusters	Image1.jpg	
	Validity Measure	**Davies-Bouldin Index**
2	4.98	2.710
3	4.556	3.632
4	7.851	**1.662**
5	**4.470**	2.937
6	4.954	2.380
7	4.819	3.100
8	8.661	3.789
9	8.199	3.982
10	46.11	4.150

Table 2b: Validity Measure and Davis-Bouldin Index for the given breast cancer Image2.jpg based on the number of clusters

Number of Clusters	Image2.jpg	
	Validity Measure	**Davies-Bouldin Index**
2	3.099	1.396
3	3.752	2.730
4	2.89	**1.209**
5	**1.946**	1.897

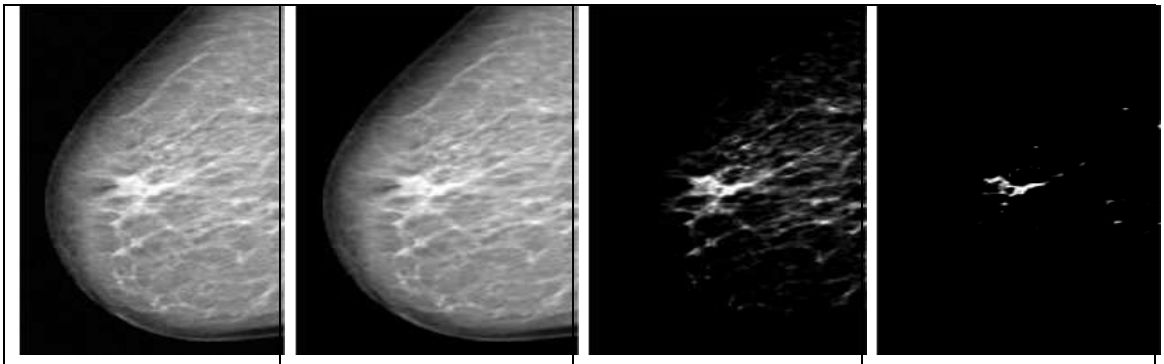
6	3.321	1.599
7	2.652	2.477
8	8.253	2.474
9	7.577	3.751
10	7.710	3.522

Table 2c: Validity Measure and Davis-Bouldin Index for the given breast cancer Image3.jpg based on the number of clusters

Number of Clusters	**Image3.jpg**	
	Validity Measure	**Davies-Bouldin Index**
2	4.126	2.854
3	6.686	2.323
4	3.833	**1.522**
5	3.487	2.533
6	**1.564**	2.987
7	3.122	2.495
8	14.63	2.788
9	12.51	2.779
10	30.28	3.947

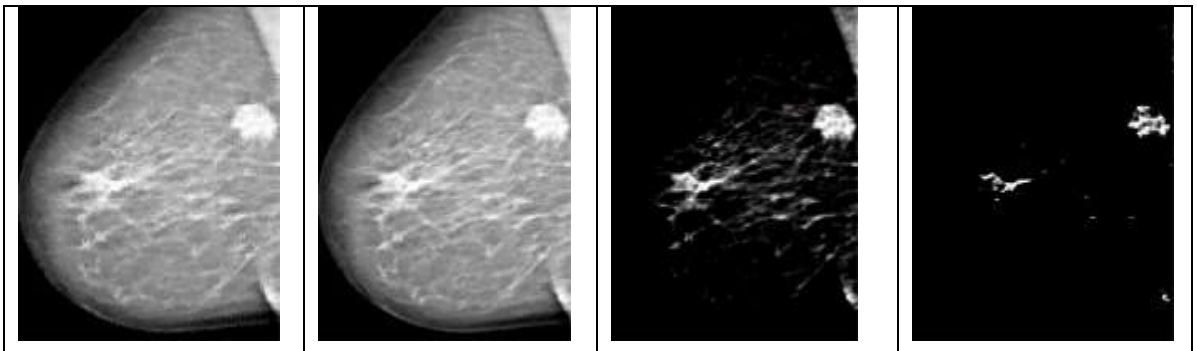
Table 2d: Validity Measure and Davis-Bouldin Index for the given breast cancer Image4.jpg based on the number of clusters

Number of Clusters	**Image4.jpg**	
	Validity Measure	**Davies-Bouldin Index**
2	**2.422**	4.947
3	108.2	4.896
4	15.69	**2.750**
5	12.86	3.221
6	18.77	4.960
7	14.68	3.331
8	13.37	4.944
9	13.79	4.667
10	16.76	4.376



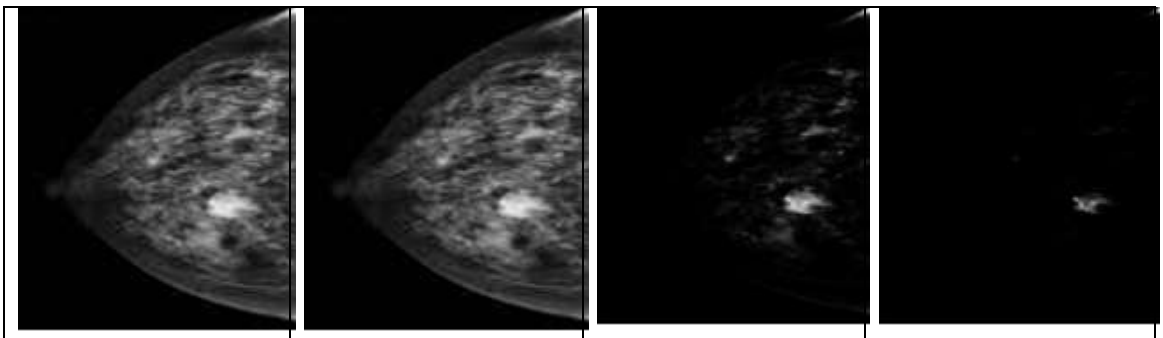
a) Original Image	b) Pre-processed Image	c) Segmented Image	d) Resultant Segmented Image

Figure 4: **Results obtained for the given Image1.jpg by Riotous Clustering and SOM Segmentation**



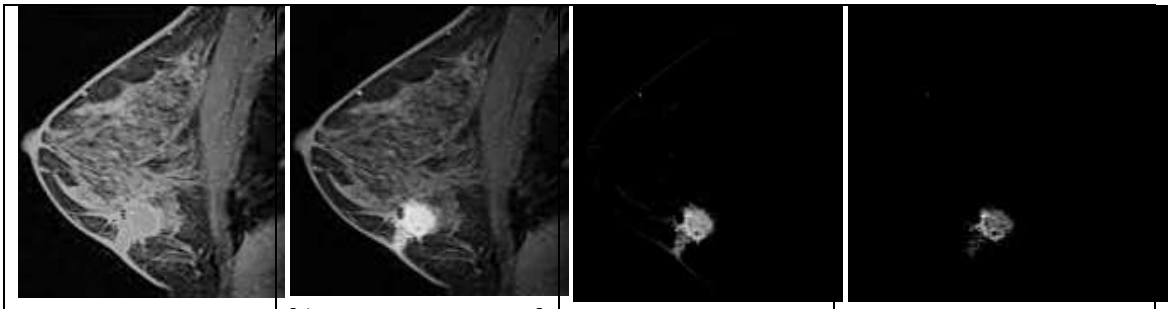
a) Original Image	b) Pre-processed Image	c) Segmented Image	d) Resultant Segmented Image

Figure 5: **Results obtained for the given Image2.jpg by Riotous Clustering and SOM Segmentation**



a) Original Image	b) Pre-processed Image	c) Segmented Image	d) Resultant Segmented Image

Figure 6: **Results obtained for the given Image3.jpg by Riotous Clustering and SOM Segmentation**



a) Original Image | b) Pre-processed Image | c) Segmented Image | d) Resultant Segmented Image

Figure 7: **Results obtained for the given Image4.jpg by Riotous Clustering and SOM Segmentation**

Above figures 4 to figure 7 presents the result obtained by proposed framework for given image1 to image 4. From the segmented image, the cancerous tissue can be obtained easily.

Table 3: **Optimal Cluster for the given breast cancer images**

Sl.No	**Image Number**	**Optimal Number of Cluster**	
		Validity Measure	Davies-Bouldin Index
1	Image1.jpg	5	2
2	Image2.jpg	5	2
3	Image3.jpg	6	2
4	Image4.jpg	2	2

From the table 3, the cluster number 2 gives the optimal solution by using Davies-Bouldin Index among the other cluster numbers for the given 4 images.

5. Conclusion

In this paper, from the proposed framework, Normalized Euclidean Distance performs clustering well and gives segmentation results as in human perception. Using this proposed work, the segmentation of the breast cancer images can be done unsupervised and automatically, by utilizing measurement of cluster validity. Davies-Bouldin Index (DBI) and Validity Measurement (VM) indexes comparatively affords distinct of optimal number of clusters. For each breast cancer images, the optimal numbers of clusters which are developed by DBI, on average are less than the results which are obtained by VM.

Reference

1. Williams, Lovoria B., et al. "Demographic, psychosocial, and behavioral associations with cancer screening among a homeless population." Public Health Nursing (2018).

2. Henriksen, Emilie L., et al. "The efficacy of using computer-aided detection (CAD) for detection of breast cancer in mammography screening: a systematic review." Acta Radiologica (2018): 0284185118770917.

3. Siddharth Singh Chouhan, Ajay Kaul, Uday Pratap Singh, “Image Segmentation Using Computational Intelligence Techniques: Review”, Archives of Computational Methods in Engineering, (2018).

4. Ahmed, Isra O., Banazier A. Ibraheem, and Zeinab A. Mustafa. "Detection of Eye Melanoma Using Artificial Neural Network." Journal of Clinical Engineering 43.1 (2018): 22-28.

5. Shukla, Nagesh, et al. "Breast cancer data analysis for survivability studies and prediction." Computer Methods and Programs in Biomedicine 155 (2018): 199-208.

6. Arora, Shaveta, Madasu Hanmandlu, and Gaurav Gupta. "Filtering impulse noise in medical images using information sets." Pattern Recognition Letters (2018).

7. Boemer, Fabian, Edward Ratner, and Amaury Lendasse. "Parameter-free image segmentation with SLIC." Neurocomputing 277 (2018): 228-236.

8. Park, Young-Seuk, et al. "Multivariate Data Analysis by Means of Self-Organizing Maps." Ecological Informatics. Springer, Cham, 2018. 251-272.

9. Kumar, Krishan, Deepti D. Shrimankar, and Navjot Singh. "Eratosthenes sieve based key-frame extraction technique for event summarization in videos." Multimedia Tools and Applications 77.6 (2018): 7383-7404.

10. Ngo, Long Thanh, Trong Hop Dang, and Witold Pedrycz. "Towards Interval-Valued Fuzzy Set–based Collaborative Fuzzy Clustering Algorithms." Pattern Recognition (2018).

11. Upendran, V., & Gopinath, R., Feature Selection based on Multicriteria Decision Making for Intrusion Detection System, International Journal of Electrical Engineering and Technology, 11(5), 217-226 (2020).

12. Upendran, V., & Gopinath, R., Optimization based Classification Technique for Intrusion Detection System, International Journal of Advanced Research in Engineering and Technology, 11(9), 1255-1262 (2020).

13. Subhashini, M., & Gopinath, R., Employee Attrition Prediction in Industry using Machine Learning Techniques, International Journal of Advanced Research in Engineering and Technology, 11(12), 3329-3341 (2020).

14. Rethinavalli, S., & Gopinath, R., Classification Approach based Sybil Node Detection in Mobile Ad Hoc Networks, International Journal of Advanced Research in Engineering and Technology, 11(12), 3348-3356 (2020).

15. Rethinavalli, S., & Gopinath, R., Botnet Attack Detection in Internet of Things using Optimization Techniques, International Journal of Electrical Engineering and Technology, 11(10), 412-420 (2020).

16. Priyadharshini, D., Poornappriya, T.S., & Gopinath, R., A fuzzy MCDM approach for measuring the business impact of employee selection, International Journal of Management (IJM), 11(7), 1769-1775 (2020).

17. Poornappriya, T.S., Gopinath, R., Rice Plant Disease Identification using Artificial Intelligence Approaches, International Journal of Electrical Engineering and Technology (IJEET), 11(10), 392-402 (2020).

18. Poornappriya, T.S., Gopinath, R., Application of Machine Learning Techniques for Improving Learning Disabilities, International Journal of Electrical Engineering and Technology (IJEET), 11(10), 403-411 (2020).