Journal of Advances in Science and Technology

Vol. IV, Issue No. VII, November-2012, ISSN 2230-9659

BREAST CANCER DETECTION USING FUZZY C-MEANS

Journal of Advances in Science and Technology
Vol. IV, Issue No. VII, November-2012, ISSN 2230-9659

Breast Cancer Detection Using Fuzzy C-Means

Arman Rasool Faridi*

Department of Computer Science, Aligarh Muslim University, Aligarh, India

Abstract – The Fuzzy C-means (FCM) algorithm is one of the most powerful algorithms used in numerous experiments that attempts to segment various types of population, and hence it can be used as a method capable of assisting physicians in early breast cancer diagnosis. In this article, a method based on the FCM algorithm is proposed for extracting clusters of various type of cancers from information gathered through electrical impedance calculation. The proposed approach aims to create clusters dependent on different characteristics such as impedance at zero frequency, among others. Once clusters have been formed, new values may be plotted on the graph, and the type of cancer a person might be suffering from can be determined based on the location of the projection.

Keywords—Breast Cancer, Fuzzy C-Means, Clustering, Segmentation, Matlab

- X -

I. INTRODUCTION

Breast cancer is most common in women over the age of forty; one possible explanation is the accumulation of cell genetic mutations over time, which can lead to cancer [1]. Mammography is still the most widely used diagnostic tool in the world, owing to its high sensitivity, which allows even microcalcification identification [2][3]. However, there are some drawbacks to this approach, such as patient discomfort caused by breast squeezing during the X-ray imaging process and poor specificity because 80 per cent of cancer cases are shown to be negative [4].

Ultrasound is the second most often employed method for distinguishing between cystic and non-cystic growths, and it has the bonus of being a painless treatment that can be used for many explorations. While extremely sensitive, magnetic resonance imaging is costly and is almost only used for specific health situations such as breast implants or possible multifocal carcinoma. These therapies are effective, but they have some drawbacks, such as painful treatments, high-cost supplies, and a lack of availability in certain developed countries for some vulnerable areas.

As a result, in recent decades, new prebiopsy diagnostic methods for global use have been introduced to reduce the number of patients who perform unnecessary biopsy procedures. Microwave imaging [5], magnetic resonance elastography [6], thermography [7], and optical mammography [8] are examples of these laboratory techniques. Owing to their poor performance, all of these methods are currently being studied and expanded upon. In recent decades, experimental techniques focused on bioimpedance calculation have piqued interest, owing to their minimally invasive nature and low cost compared to those described above [9].

Electrical impedance tomography [10], electrical impedance spectroscopy [11], and electrical impedance mammography [12] are examples of techniques developed by this research. The most fundamental of these methods is bioimpedance tissue classification, which has demonstrated that the conductivity of carcinoma can be 20 to 40 times greater than that of stable breast tissue in vitro studies of breast tissues [13]. While there has been very little in vivo testing, the current trials have reported some improvement in the experimentally acquired findings, which encourages more bioimpedance research into breast carcinoma. Until conducting in vivo experiments, newly proposed approaches for breast carcinoma detection are tested on testing models (phantoms) as the first experimental step. Within such research models, biological tissues of interest in the sample (normal, healthy, and malignant) are simulated. For such breast phantoms, different manufacturing materials and textures have been documented[14] [15][16].

The majority of studies that have used electrical impedance to diagnose breast cancer have attempted to recreate anatomical photographs of the mammary tissues from bioimpedance measurements in order to pinpoint the site of a potentially cancerous tumour in the preclinical stage (diameter less than or equal to 1 cm) [17]. Surface electrodes, which are less intrusive than needle electrodes, are used in both of these studies, but their contact region is dependent on the contact impedance, and current density supplied. Microelectrodes, for example, have a higher contact impedance than electrodes with a diameter of 1 cm, and their current density is equal to their surface area in both instances. The most common number of electrodes in a ring configuration is 16 or 32, but 128 or just four have also been recorded. In some of these methodologies, a large number of measuring electrodes are used to improve high precision, and in others, limited measuring electrodes are used to increase the current density inside the tissue[18].

Soft clustering (also referred to as fuzzy clustering or soft k-means) is a form of clustering in which each data point is allocated to several clusters. Clustering, also known as cluster analysis, is a technique of grouping data points into clusters in such a way that events in the same cluster are as similar as possible, whereas artifacts in different clusters are as dissimilar as possible. Clusters are identified using similarity checks. Similarity tests include things like width, connectivity, and strength. Different similarity metrics may be chosen depending on the data or the program. [19].

In this paper, the fuzzy c-means technique has been used to make the cluster of different type of cancers based on the Impedance measurement. Firstly, the background of fuzzy c-means and explanation of the dataset is done. Next, how the experiment is performed in MATLAB is explained, and then the results are discussed along with the conclusion.

II. BACKGROUND

A. Fuzzy c-means

Clustering, also known as cluster analysis, is the process of grouping data points into clusters in such a way that events in the same cluster are as identical as possible, while things in different clusters are as different as possible. Similarity tests are used to identify clusters. Distance, connectivity, and strength are examples of similarity tests. Depending on the data or the application, different similarity metrics may be selected [19].

A clustering process called FCM (fuzzy c-means) requires a single piece of data to belong to two or more clusters. This approach is widely used in pattern recognition. It is based on the following objective function's minimisation:

$$O_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m} \parallel x_i - c_j \parallel^2 , \quad where\ 1 \le m < \infty$$

x_i is the ith of d-dimensional measured data, c_j is the d-dimension centre of the cluster, and ||*|| is any norm representing the resemblance between any measured data and the centre, where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension centre of the cluster.

Fuzzy partitioning is accomplished by iteratively optimising the objective function shown above, with membership u_{ij} and cluster centres c_j modified as follows:

$$u_{ij} = \frac{1}{\sum_{k=1}^{C}\left(\frac{\parallel x_i - c_j \parallel}{\parallel x_i - c_k \parallel}\right)^{\frac{2}{m-1}}} \quad where\ c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{1}^{N} u_{ij}^{m}}$$

This iteration will end when $\max_{ij}\{|u_{ij}^{k+1} - u_{ij}^{k}|\} < \varepsilon$, where ε is a criterion for termination and k is the number of iterations. This method converges to an O_m saddle point or local minimum.

The steps in the algorithm are as follows:

1. Initialize matrix U=[u_{ij}], $U^{(0)}$

2. At nth-step: calculate the centers vectors C(n)=[cj] with U(n)

$$c_j = \frac{\sum_{i=1}^{N} u_{ij}^{m} . x_i}{\sum_{1}^{N} u_{ij}^{m}}$$

3. Update $U^{(n)}$, $U^{(n+1)}$

$$u_{ij} = \frac{1}{\sum_{n=1}^{C}\left(\frac{\parallel x_i - c_j \parallel}{\parallel x_i - c_n \parallel}\right)^{\frac{2}{m-1}}}$$

4. If $\max_{ij}\{|u_{ij}^{n+1} - u_{ij}^{n}|\} < \varepsilon$ i.e. stopping criterion is met, then STOP; otherwise, return to step 2.

B. Description of the dataset

The dataset used includes six diseases and nine elements. Fibro-adenoma, Carcinoma, Glandular, Connective, Mastopathy and Adipose are six disorders. Carcinomas are cancers or malignancies that start in epithelial cells, which make up the skin and the tissues that line internal organs and structures. Breast, liver, kidney, and colon carcinomas are among the most prevalent cancers.

Fibroadenomas are noncancerous breast lumps that mostly concern women between the ages of 15 and 35. A fibroadenoma has a distinct shape which can be sturdy, smooth, rubbery, or rough to the touch. It may feel like a marble in the breast, slipping effortlessly under the skin when tested, and is usually painless. Fibroadenomas come in a variety of sizes and can grow or shrink on their own.

Mastopathy is a benign alteration in the glandular tissue of the breast that is caused by hormones. Symptoms are most pronounced only before and after menstruation. Mastopathy refers to a variety of benign alterations in the mammary glands, including nodules, swellings, and cysts.

Cancer starts in the glandular (secretory) cells of the body. The lobes, which produce milk, and the ducts, which transport milk to the nipple, are also known as glandular tissue. Fibrous and glandular tissue are combined to form fibroglandular tissue.

Breast sarcomas are a form of breast cancer that is very rare. Sarcomas account for less than one per cent of all breast cancers diagnosed. Breast sarcomas start in the connective tissue that protects the ducts and lobules of the breast, unlike most breast cancers that start in the milk ducts.

Breast adipose tissue's fundamental purpose is to retain stored energy and release it when the body requires it. Breast adipose tissue, on the other hand, plays an essential part in breast growth and maturation. It also helps the growth and advancement of breast cancer because it is a rich energy supply. Breast adipose tissue secretes a number of growth factors that cancer cells use to stay alive.

In this study, nine features as a result of electrical impedance have been taken. When a voltage is applied to a circuit, the electrical impedance is the amount of resistance it provides to a current. The ratio of the complex representation of the sinusoidal voltage between its terminals to the complex representation of the current flowing through it is the impedance of a two-terminal circuit part. In addition, the frequency of a sinusoidal voltage determines it. Impedance extends the concept of resistance to alternating current (AC) circuits which has both magnitude and phase, unlike resistance, which only has magnitude. The following frequencies were used to calculate impedance: 15.625, 31.25, 62.5, 125, 250, 500, and 1000 kHz [20].

- Impedivity (in ohm) at zero frequency
- phase angle at 500 kHz
- the high-frequency slope of phase angle
- area under spectrum
- impedance distance between spectral ends
- maximum of the spectrum
- area normalised by DA
- length of the spectral curve
- distance between I0 and the fundamental part of the maximum frequency point

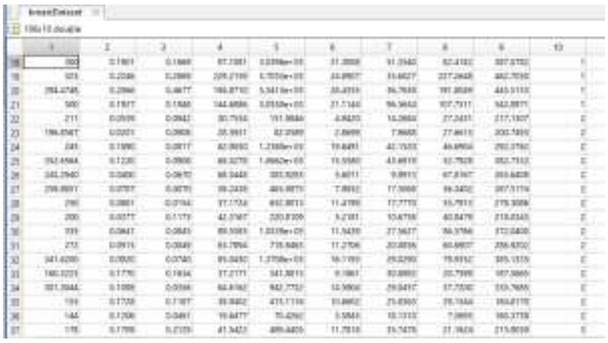
III. EXPERIMENTAL SETUP

The results of the impedance are not calculated, but it is available as standard for downloading. It consists of an excel file in which the first sheet contains the details of the diseases. This study included fourteen records for each disease containing all the features. Original data contains more than fourteen entries, and it was in the format as shown in table 1.

Table 1: Fields in the original data

S. No.	Field
1	Case # (Serial Number)
2	Class (Disease)
3	I0 (Impedivity (ohm) at zero frequency)
4	PA500 (phase angle at 500 kHz)
5	HFS (the high-frequency slope of phase angle)
6	DA (impedance distance between spectral ends)
7	Area (area under spectrum)
8	A/DA (area normalised by DA)
9	Max IP (maximum of the spectrum)
10	DR (distance between I0 and the fundamental part of the maximum frequency point)
11	P (length of the spectral curve)

From the original data, the Case field was removed. The Class field, which was in text form, was converted to the numerical field, and for the same field, same numerical value has been given. By this, data can be categorized. Sample data is shown in figure 1.



Sample data used to create clusters

First, the file name is set, and then the CSV file is read. The code has been implemented in Matlab.

```
%Setting file name

filename='BreastTissue.csv';

%reading csv by file name

breastDataset = csvread(filename);
```

Next, from the data, we will get the index for different diseases based on the last field.

```
carIndex = breastDataset(:,10)==1;

fadIndex = breastDataset(:,10)==2;

masIndex = breastDataset(:,10)==3;

glaIndex = breastDataset(:,10)==4;

conIndex = breastDataset(:,10)==5;

adiIndex = breastDataset(:,10)==6;
```

Once the indexes are obtained the last field will be removed and the data is normalized.

```
breastDataset=breastDataset(:,1:9);

normalizedData = bsxfun(@minus, breastDataset, mean(breastDataset));

breastDataset = bsxfun(@rdivide, normalizedData, std(breastDataset));
```

Next the separation of data fields is done using the indexes that were found earlier :

```
car = breastDataset(carIndex,:);

fad = breastDataset(fadIndex,:);

mas = breastDataset(masIndex,:);

gla = breastDataset(glaIndex,:);

con = breastDataset(conIndex,:);

adi = breastDataset(adiIndex,:);
```

Next removal of extra rows is done and only fourteen rows are selected for all the diseases :

```
car = car(1:14,:);

fad = fad(1:14,:);

mas = mas(1:14,:);

gla = gla(1:14,:);

con = con(1:14,:);

adi = adi(1:14,:);
```

Next, as separation is done then we can now plot the graphs between different features, but before that, we need to create characteristics and create combinations between various characteristics:

```
Characteristics =

{'I0','PA500','HFS','DA','Area','A/DA','Max IP','DR','P'};

[~,s]=size(Characteristics);

totalCount=s*(s-1)/2;

count=1;

pairs=zeros(totalCount,2);

for c = 1:(s-1)

for r = c+1:s

pairs(count,1)=c;

pairs(count,2)=r;

count=count+1;

end

end
```

So for total pairs between different features is equal to

$$N_p = N_f * \frac{(N_f - 1)}{2}$$

Where N_p = Number of pairs and N_f = Number of features

Once the pairs are created then we will create the subplots for each pair :

```
for i = 1:totalCount

x = pairs(i,1);

y = pairs(i,2);

subplot(6,6,i);

plot([car(:,x) fad(:,x) mas(:,x) gla(:,x) con(:,x) adi(:,x)],...

[car(:,y) fad(:,y) mas(:,y) gla(:,y) con(:,y) adi(:,y)], '.')

xlabel(Characteristics{x})

ylabel(Characteristics{y})

end
```

Now, these graphs contain data points between different features. And now, the fuzzy c means algorithm will be used and the centres for each cluster is found out. Before calling the fuzzy c means algorithm, we have to set various parameters of this algorithm.

```
Nc = 6;

M = 2.0;
```

```
maxIter = 100;

minImprove = 1e-6;

clusteringOptions = [M maxIter minImprove true];
```

Here Nc means the number of clusters that we want, M is the exponent for the matrix, maxIter means the maximum number of iterations, and minImprove means the minimum amount of improvement that we need. Finally, we are setting the variables in one vector, which consist of all the clustering options.

After that fuzzy c means algorithm is called and centers are collected in a variable:

```
[centers,U] = fcm(breastDataset,Nc,clusteringOptions);
```

Now, as we get the centres, then these are plotted in the subplots earlier so that we can know whether the centres are appropriate or not.

```
for i = 1:totalCount

subplot(6,6,i);

for j = 1:Nc

x = pairs(i,1);

y = pairs(i,2);

text(centers(j,x),centers(j,y),int2str(j),'FontWeight','bold'
);

end

end
```

IV. RESULTS & CONCLUSION

Sample results are shown in figures added in appendix 1. These figures contain plots for all the combinations that can be formed using features or characteristics taken from the test. Here different colour shows different clusters and as we can see centres are created using fuzzy c means algorithm are around the different colours or diseases. Some results are more related than others, like in DA vs I0, DR vs I0, P vs I0, Area vs DA and DR vs DA, it is difficult to determine the clusters. But on all remaining relations, it is easy to find the cluster and to know where the new point will belong to.

So this can be a way to know about the disease without any further test once we give details about the features and then the new point is plotted and by that, we can know what can be possible diseases. The point which is closer to a centre means that disease is more probable. In future, an application can be made where just take the features can be taken and, based on that, present the probability of the disease.

V. REFERENCES

[1] "SEER Cancer Statistics Review 1975-2006- Previous Version - SEER Cancer Statistics," SEER. .

[2] I. Damjanov (2008). “ENZINGER AND WEISS'S SOFT TISSUE TUMORS, 5TH EDITION," Shock, vol. 30, no. 6.

[3] A. H. Israyelyan (2003). "The Development of Molecular Diagnostics for Breast Cancer," p. 80.

[4] Y. Zou and Z. Guo (2003). “A review of electrical impedance techniques for breast cancer detection.," Med. Eng. Phys., vol. 25, no. 2, pp. 79–90.

[5] E. C. Fear, S. C. Hagness, P. M. Meaney, M. Okoniewski, and M. A. Stuchly (2002). Enhancing breast tumor detection with near-field imaging," IEEE Microw. Mag., vol. 3, no. 1, pp. 48–56.

[6] E. E. W. Van Houten, M. M. Doyley, F. E. Kennedy, J. B. Weaver, and K. D. Paulsen (2003). "Initial in vivo experience with steady-state subzone-based MR elastography of the human breast.," J. Magn. Reson. Imaging, vol. 17, no. 1, pp. 72–85.

[7] E. Y.-K. Ng and S.-C. Fok (2003). “A framework for early discovery of breast tumor using thermography with artificial neural network," The breast journal, vol. 9, no. 4. United States, pp. 341–343.

[8] D. Grosenick et. al. (2003). "Time-domain optical mammography: initial clinical results on detection and characterisation of breast tumors," Appl. Opt., vol. 42, no. 16, pp. 3170–3186.

[9] E. Y. K. Ng, S. V. Sree, K. H. Ng, and G. Kaw (2008). "The Use of Tissue Electrical Characteristics for Breast Cancer Detection: A Perspective Review," Technol. Cancer Res. Treat., vol. 7, no. 4, pp. 295–308.

[10] T. E. Kerner, K. D. Paulsen, A. Hartov, S. K. Soho, and S. P. Poplack (2002). "Electrical impedance spectroscopy of the breast: clinical imaging results in 26 subjects," IEEE Trans. Med. Imaging, vol. 21, no. 6, pp. 638–645.

[11] B. Singh, C. W. Smith, and R. Hughes (1979). "In vivo dielectric spectrometer," Med. Biol. Eng. Comput., vol. 17, no. 1, pp. 45–60.

[12] O. V Trokhanova, M. B. Okhapkin, and A. V Korjenevsky (2008). "Dual-frequency electrical impedance mammography for the diagnosis of non-malignant breast disease.," Physiol. Meas., vol. 29, no. 6, pp. S331-44.

[13] K. R. Foster and H. P. Schwan (1989). "Dielectric properties of tissues and biological materials: a critical review.," Crit. Rev. Biomed. Eng., vol. 17, no. 1, pp. 25–104.

[14] T. Morimoto et. al. (1990). "Measurement of the electrical bio-impedance of breast tumors.," Eur. Surg. Res. Eur. Chir. Forschung. Rech. Chir. Eur., vol. 22, no. 2, pp. 86–92.

[15] T. Morimoto et. al. (1993). "A study of the electrical bio-impedance of tumors.," J. Investig. Surg. Off. J. Acad. Surg. Res., vol. 6, no. 1, pp. 25–32.

[16] A. Stojadinovic et. al. (2005). "Electrical impedance scanning for the early detection of breast cancer in young women: preliminary results of a multicenter prospective clinical trial.," J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol., vol. 23, no. 12, pp. 2703–2715.

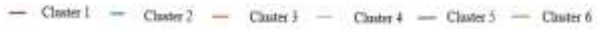
[17] G. A. Ybarra et. al. (2007). "Breast imaging using electrical impedance tomography (EIT)," Emerg. Technol. Breast Imaging Mammography, Am. Sci. Publ., pp. 1–16.

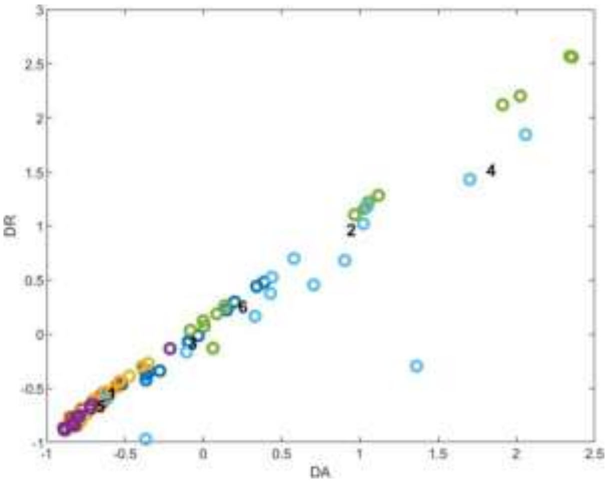
[18] J. Prado, C. Margo, M. Kouider, and M. Nadi (2005). "Impedance of electrolytes using microelectrodes coplanar," in Proceeding of COMSOL Multiphysics Conference, pp. 241–245.

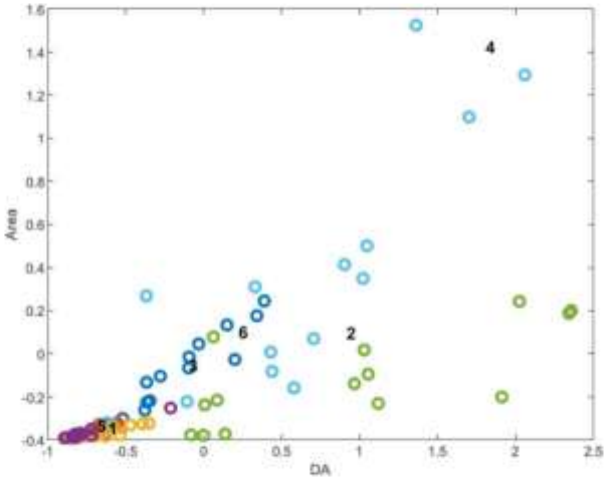
[19] J. C. Bezdek, R. Ehrlich, and W. Full (1984). "FCM: The fuzzy c-means clustering algorithm," Comput. Geosci., vol. 10, no. 2–3, pp. 191–203.

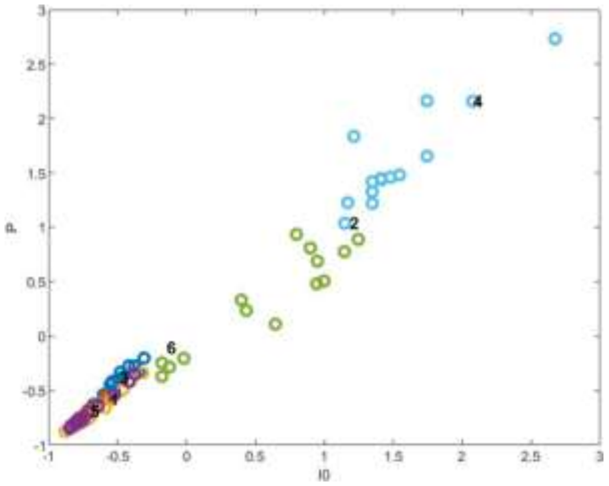
[20] "Datasets Breast Cancer." URL: http://archive.ics.uci.edu/ml/datasets.php.

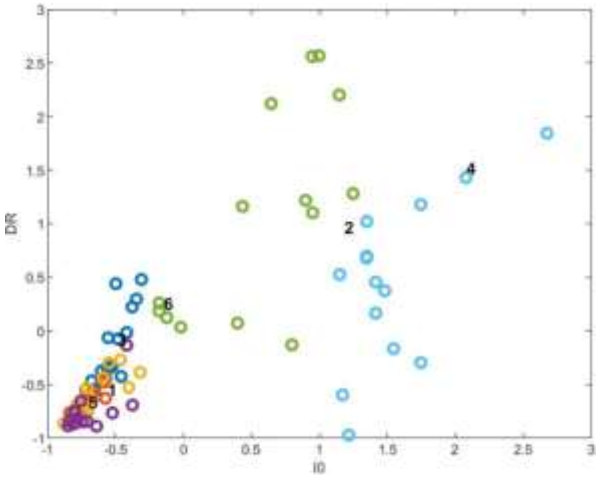
APPENDIX 1

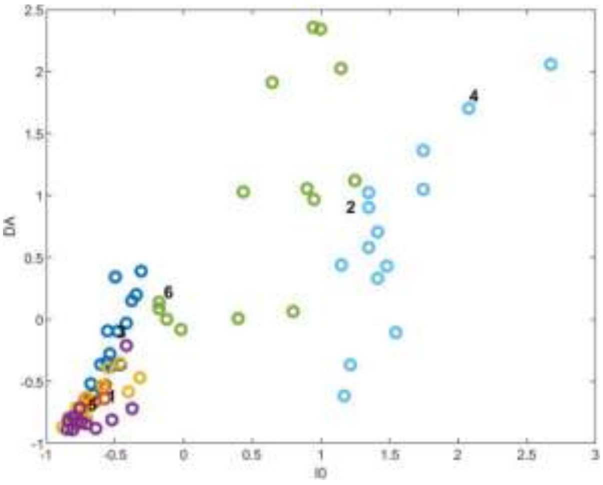


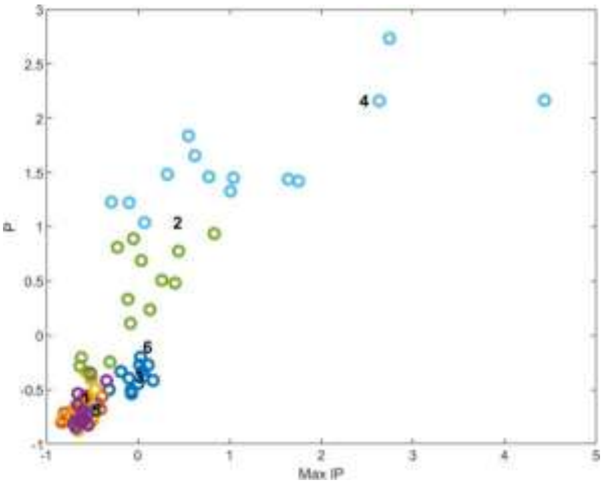












Corresponding Author

Arman Rasool Faridi*

Department of Computer Science, Aligarh Muslim University, Aligarh, India

arman.faridi@gmail.com