

ANALYSIS OF DNA MICROARRAY IMAGES USING GRIDDING AND SEGMENTATION METHOD

*Ulli Moulali, **Dr. Syed Umar

**Research Scholar, Department of Computer Science, Faculty of Computing & Information Technology, Himalayan University, Itanagar*

***Supervisor, Department of Computer Science, Faculty of Computing & Information Technology (Computer Science), Himalayan University, Itanagar*

ABSTRACT:

This paper mainly addresses meshing and segmentation approaches for microarray image analysis. The word "grid" refers to partitioning an image into subgrids of dots and then separating them into point detection. Most of the current approaches rely on input factors such as the number of rows / columns, the number of points in each row / column, the size of the subarrays, etc. This article presents a completely automated mesh generating technique. This may delete any initialized parameter without any user interaction. In the segmentation stage, clustering methods are utilized since they do not concern the size and form of the spots, do not rely on the initial state of the pixels, and do not need post-processing. In this article, a novel approach is provided to predict the initial parameters (centroid and number of clusters) needed by any clustering algorithm. Qualitative and quantitative examination demonstrates that the method can conduct grid processing on microarray pictures effectively, and increases the performance of the clustering approach.

Keywords: *Microarray image, Image segmentation, Mathematical morphology, clustering algorithms*

1. INTRODUCTION:

Implementing a conditional convolution filter [7] followed by transforming the picture into a binary image using the histogram transformation function, DNA microarray gridding may be achieved. Using sub-block histogram equalization, this method is possible to pinpoint each location. Mesh generation is totally automated in this Paper. Delete any initialised parameter without user input. During the segmentation stage, clustering algorithms have been used since they do not need post-processing and are not concerned with the size or shape of spots. Any clustering algorithm's initial parameters (centroid and cluster count) may be predicted using a new technique shown in this article. A combination of qualitative and quantitative research shows that the technique is successful at performing process of gridding on micro array images and enhances the clustering approach's effectiveness. Genome expression levels at thousands of sites at the same time may be assessed using microarray technology Image of hybrid microarray slide captured by sensor with two unique wavelengths, Cy3 and Cy5, is result of microarray experiment. There are three steps in the microarray image analysis process: gridded images are segmented, and then quantified. There are two independent phases to the gridding process.

Sub-arrays (also called sub-grids) may be divided into gene point regions (also known as point detection), as illustrated in Figure 1. Spot and background pixels in each spot area are segmented into smaller subgroups via the process of segmentation. Genome expression may be quantified by

measuring the logarithmic ratio of a point's green and red intensity values, which is a statistical analysis's result [2].

Foreground pixels (spot regions) and background pixels have different intensities, and the gene expression value is dependent on these differences. For microarray image analysis, the majority of currently available techniques are semi-automatic, that implies that operator intervention is as necessary to establish the parameters and then run the algorithm of gridding [3,4]. A completely automated mesh generating method is presented in this article. Microarray technology is extensively employed in many fields, including genetics, illness diagnostics, drug research, and pharmacology [4, 5].

In this paper, the second part presents a grid generation method, the technique for computing the clustering algorithm's needed parameters is shown in the third section, the technique for clustering is explained in detail in the fourth part, followed by a discussion of the findings and recommendations drawn from the experiments.

2. LITERATURE

The use of DNA microarrays allows researchers to examine the expression of many genes throughout the whole genome, revealing new details about the genome's activities. With so many genes and biological information, it might be difficult to identify a person's specific genetic profile. It is very difficult to extract useful information. Microarray technology is the best method for studying gene expression patterns, according to M. T. Miller et al. 2003 [31]. Microarray data processing was also described by N. Giannakeas et al. clustering genes with comparable expression patterns is the first stage in the DNA microarray analysis process.

Lin et al. used microarrays to uncover gene regulatory networks and cellular processes; P. O'Neill et al. used functional genomics to detect gene expression patterns in various samples. In 2003 [32], Giacomini et al. used microarray technology to identify cancer subtypes utilizing multi-class cancer diagnosis.

DNA microarray data may be classified using clustering and regression approaches, according to H. Ling et al. 2007 [33]. Other techniques include: According to Li et al., a Bayesian statistical framework with descriptive analysis, Image metrics, and discriminant analysis are the best ways of analysing variation in microarrays. Other statistical models for big microarray data sets have been suggested by Zhao et al.

These include single and full link hierarchical clustering; the K family of clustering algorithms; optimization-based techniques; fuzziness and quality thresholds; SOTA, simulated annealing; as well as information-based clustering and the self-organizing map and SOTA [29].

For microarray data analysis, several methods have been suggested. As part of their analysis of gene networks in lung cancer, Eisen et al. used the Backward Chaining rule. In the same way, there are a variety of ways to analyse cancer data. DNA microarray data normalization has been studied by Yeung et al. 2003 [35]. Gene Cluster 2.0 and microarray explorer tools for microarray data from the

mammary gland are well-known software methods. Since genes with comparable functions are expressed in a similar way, clustering may be utilised to determine the functions of less-known genes.

Clustering gene expression data may be accomplished using a variety of techniques, each with its own set of drawbacks. Thus, regardless of the approach utilized, the quality of the clusters may be judged by a variety of ways. Using the number of cluster labels for fresh data, Jiang et al. 2003 [36] established an approach for assessing cluster quality that was extended by Yeung et al.

To verify the biological coherence of each cluster, Welsh et al. accessed either the Martinsried Institute of Proteins Sciences or the Gene Ontology databases. Any particular cellular activity or scientific research relies on a limited number of genes, making the Gene Ontology Consortium and finding and isolating outlier genes all the more critical. Similarly, genes with identical expression patterns may not execute the same tasks, as is the case with genes. Such difficulties need the use of knowledge-based clustering algorithms, such as those presented by Chang et al 2019 [37].

Using microarray image data, Jain [25] suggested a method for entirely automated quantification. The X and Y signal intensity peaks are utilised to determine the raster picture spot space and sub-array spacing. The foreground and background pixels are distinguished using a histogram that spans a squared region centred on the point of interest. When looking at the spot area, the foreground pixels have a high end of distribution, but the background pixels have a low end of distribution, as seen in the figure. There are a huge number of spots required by the technique, and the system is not resilient to misalignment of various grids.

Y.Wang [18] demonstrated how to grid microarray pictures with pinpoint accuracy. The mechanism is composed of three stages: first, the pre-processing stage, during which the global variables for gridding are calculated; second, the gridding stage; and third, the post-processing stage. Second, rotation detection, in which rotated sub-arrays are identified by calculating local gridding variables, which is a variation on the previous technique. This step is followed by a refining stage based on the local information of each sub-array, which is the third stage of local gridding. The technique is vulnerable to contamination and has a high number of missing spots, which makes it difficult to use.

Shuqing Zhao suggested a method for analyzing microarray images that relied on mathematical morphology. This paper proposes an improved gridding approach that is depended on mathematical morphology and is characterized by filtering out block noise and filtering projection plots, among other things. While gridding and spotting, certain parameters concerning the sub-array and spots are necessary through this may either be setup in advance or collected from a database during the operation. Sub-grids aren't automatically determined, and the technique is vulnerable to data-dependent parameter fluctuations.

Deepa.J [43] presented automated gridding of DNA microarray pictures using the optimal sub-image, which was implemented in the software. The technique is depended on the best sub-image selection, and the parameters for gridding are determined based on the intensity projection profile of the best sub-image that was selected. It is possible to find the ideal sub-image by looking for a block with the highest mean intensity. This approach works effectively when dealing with photos that have bright areas. A three-stage procedure is used to implement the mechanism: 1) preprocessing, 2) global parameter estimate for detecting sub-arrays, and 3) local parameter estimation for locating spots

inside the sub-array. Because of the noise pollution, the accuracy of this procedure reduces as the amount of noise contamination grows.

In the case of organisms that are not adequately annotated, however, this would be considered a failure. This necessitates the development of a new technique in which genes with similar activities are given a shared prior probability rather than genes with diverse roles. The grouping of genes with uncertain functions cannot be done in a fair manner using this technique. Detweiler et al. 2003 [33] evaluated an innovative strategy in which knowledge-based clustering was utilised to selectively trim the hierarchical tree in order to be compatible with the existing functional annotations. Other clustering approaches, such as those published by W-B. Tao et al., are superior than this method. Marchell et al. published a similar study in which the distances between genes in the same cluster are not changed, but the distances are refined in a second refinement phase to integrate information about GO annotations.

J. Harikiran et al. Journal of Hyperspectral Imaging and Image Segmentation, Volume 34, Number 34, 2017, proposes a framework for hyperspectral image segmentation that is based on a clustering method [62]. The framework for segmenting a hyper spectral data set consists of four phases that must be completed. Filtering is performed at the first step in order to remove noise from picture bands. Two stages are involved: dimensionality reduction methods and the removal of bands that contain little information or are redundant. The informative bands that were picked in the second step are fused into a single picture in the third stage, which is accomplished via the use of a hierarchical fusion approach. There is an organized hierarchy of images formed in the fusion reaction, with each group receiving an equal amount of images. This process results in a collection of photos with a great deal of variation in information, which lowers the quality of the fused image.

In spite of the large number of clustering algorithms available for microarray data processing, each approach has its own set of flaws that must be addressed. As a result, a more accurate means of improving the accuracy of microarray technology is necessary.

3. MICROARRAY GRIDDING:

When it comes to microarray image analysis, the grid is the most important step. As microarray images are analysed in a sequential fashion, the segmentation and quantification processes benefit from using a fine mesh. Each unique gene spot on a picture of a microarray is captured by subdividing the image into several spot areas. These spot regions are further split into even smaller sub regions. The method begins with a global meshing (sub-meshing) phase and concludes with a local meshing step (point detection). A region with a single spot and a backdrop is created after the mesh is finished. Manual and semi-automatic meshing algorithms are two types of meshing algorithms that are currently available. When creating a mesh by hand, in order to generate a mesh, the user must input all the necessary parameters, such as how many sub-arrays are needed and how many points are needed, among other things. When using semi-automatic software, the number of rows and columns, as well as the amount of points per row and column, may be set by the user. This Paper explains how to split the grid using the x and y direction contours of the microchip image. Algorithms for detecting sub-grids and points are shown in Figures 6.1 and 6.2, respectively.

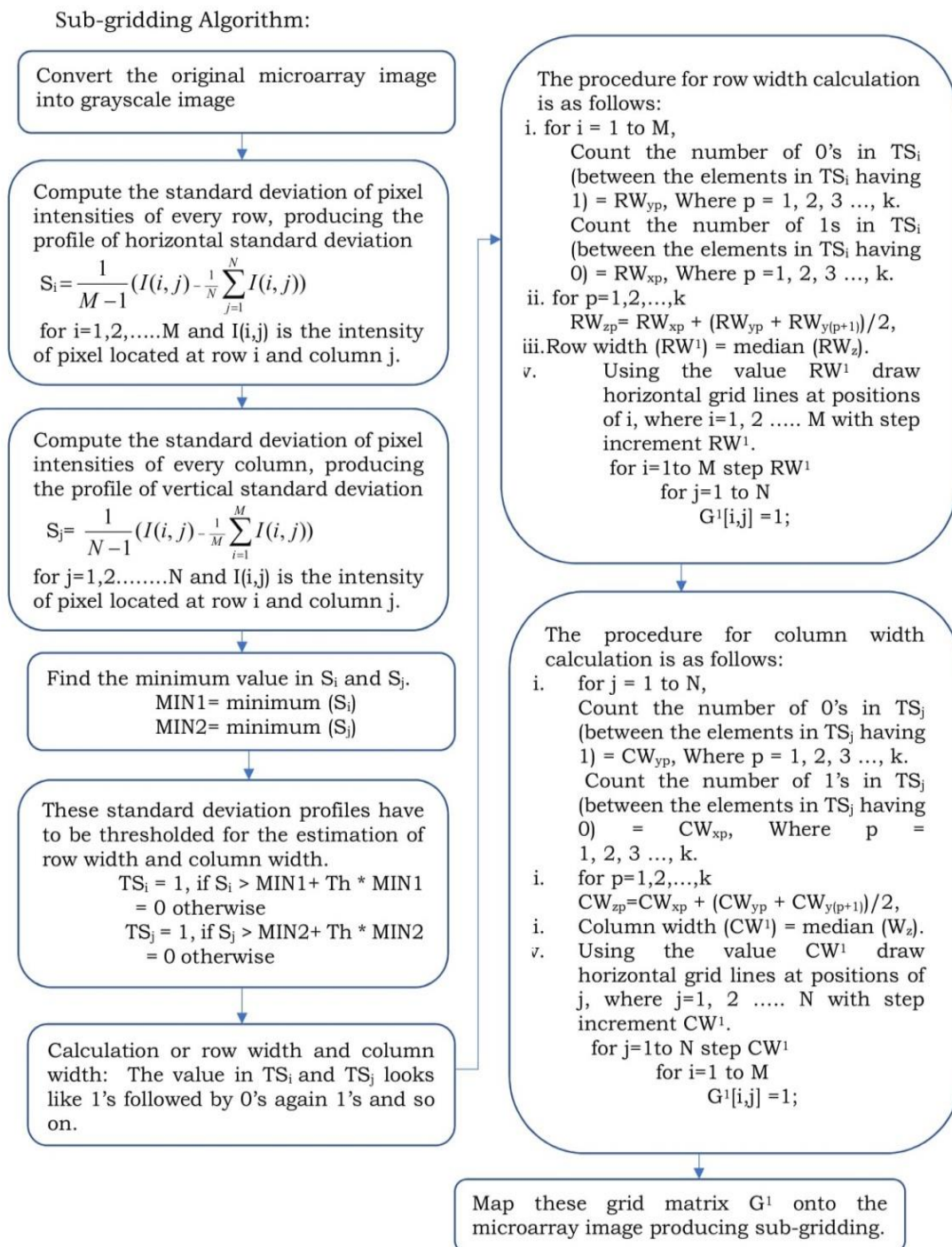


Figure 6.1: Sub – Gridding Algorithm

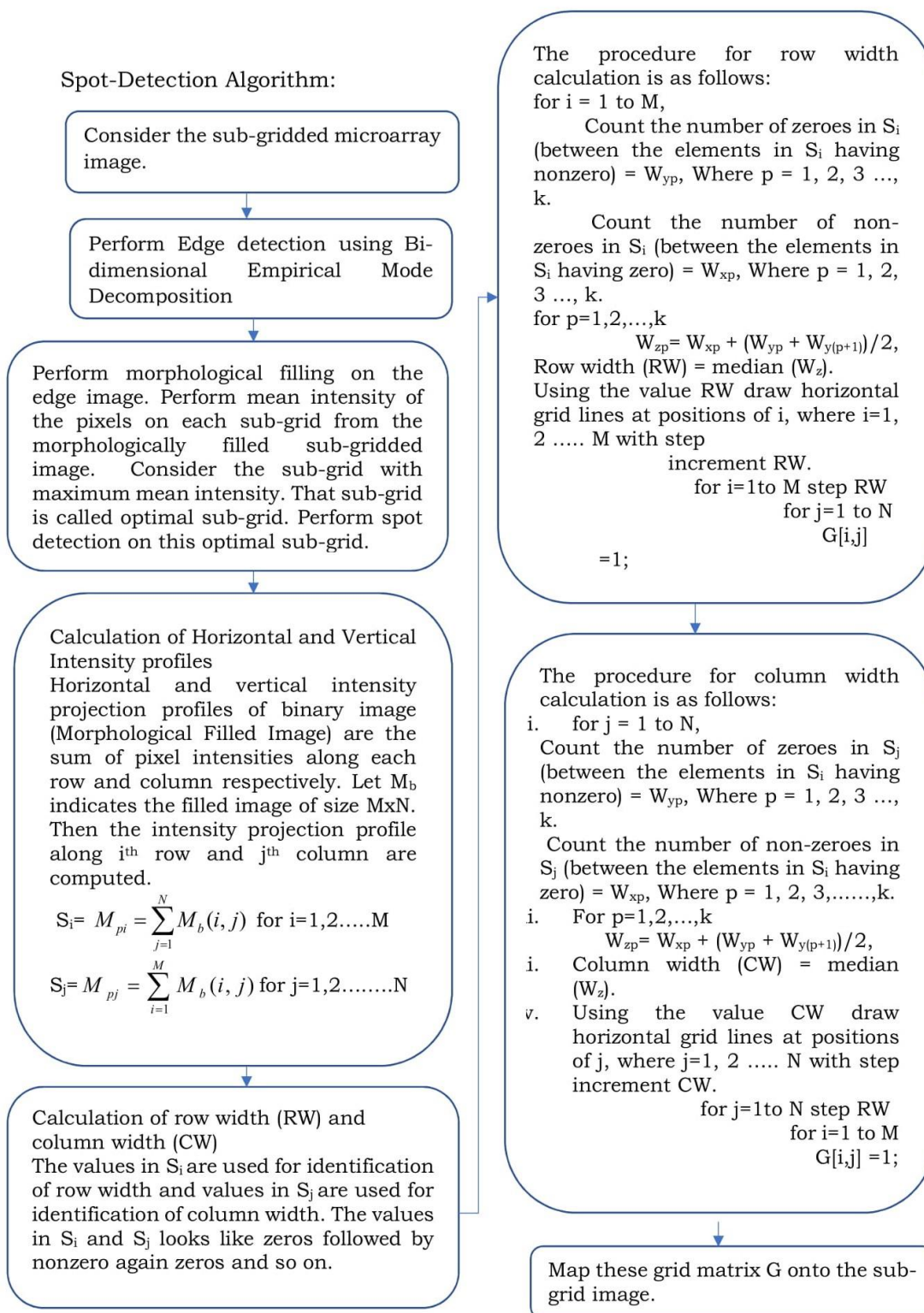


Figure 3.1: Spot Detection Algorithm

MICROARRAY SEGMENTATION:

When working with microarray pictures, the procedure of segmentation is used to divide a subnet into foreground and background areas. It is possible to think of picture segmentation as the split of an image into distinct parts. This composition has a foreground and a spot region. Based on a segmented picture spot at this area, levels of gene expression are determined. Since the perceived effort in this place is various and the types and sizes of spots are different, it is difficult to distinguish between them. A variety of statistical segmentation approaches have been presented for the segmentation of microarray pictures. [6] A method called shape-based circular segmentation is used to detect individual points in an image. There are, however, a few differences in the form of the dots. Seed pixels are selected from both the spot and the background areas, and regions are extracted by applying certain predetermined criteria to the seed pixels from both areas. The selection of a seed pixel is a time-consuming process. The thresholding method [6] estimates a suitable threshold for an image by analysing the histogram of the image, after which the image is divided into two regions. The Mann Whitney test is used to make this determination. Morphology-based segmentation [7] is a technique for segmenting images that makes use of morphological operations such as hit or miss transforms. The selection of the mask that will be used for morphological operations is critical in this case. The support vector machine algorithm [8] is used to segment the image in the context of supervised learning-based segmentation. In this Paper, image segmentation is accomplished through the use of clustering algorithms. Unlike previous segmentation approaches, these algorithms don't depend on the spot's shape or size to determine their efficacy. Neither a mask nor seed pixels are required for this approach to function properly. Every clustering procedure is influenced by the number of clusters and the starting values of centroids in the data set. In the event that these values can be approximated, the number of iterations needed by any method to segment a picture may be reduced to a minimum. In order to segment microarray images, the needed number of centroids and clusters is two, since every method separates the picture into two regions: the background area and the spot area. For segmenting the picture into two sections, Existing methods employed minimum and maximum values for background and front, respectively, but this was not the case. Another paper explains in detail how to use empirical mode decomposition to estimate centroids and the number of clusters. As seen in Figure 6.3, the number of clusters in a dataset may be estimated using this block diagram (Figure 6.3). Data are grouped using clustering approaches including k-mean, moving k-mean, and FCM in this Paper.

Algorithm for Estimation of Centroids and Number of Clusters using EMD:

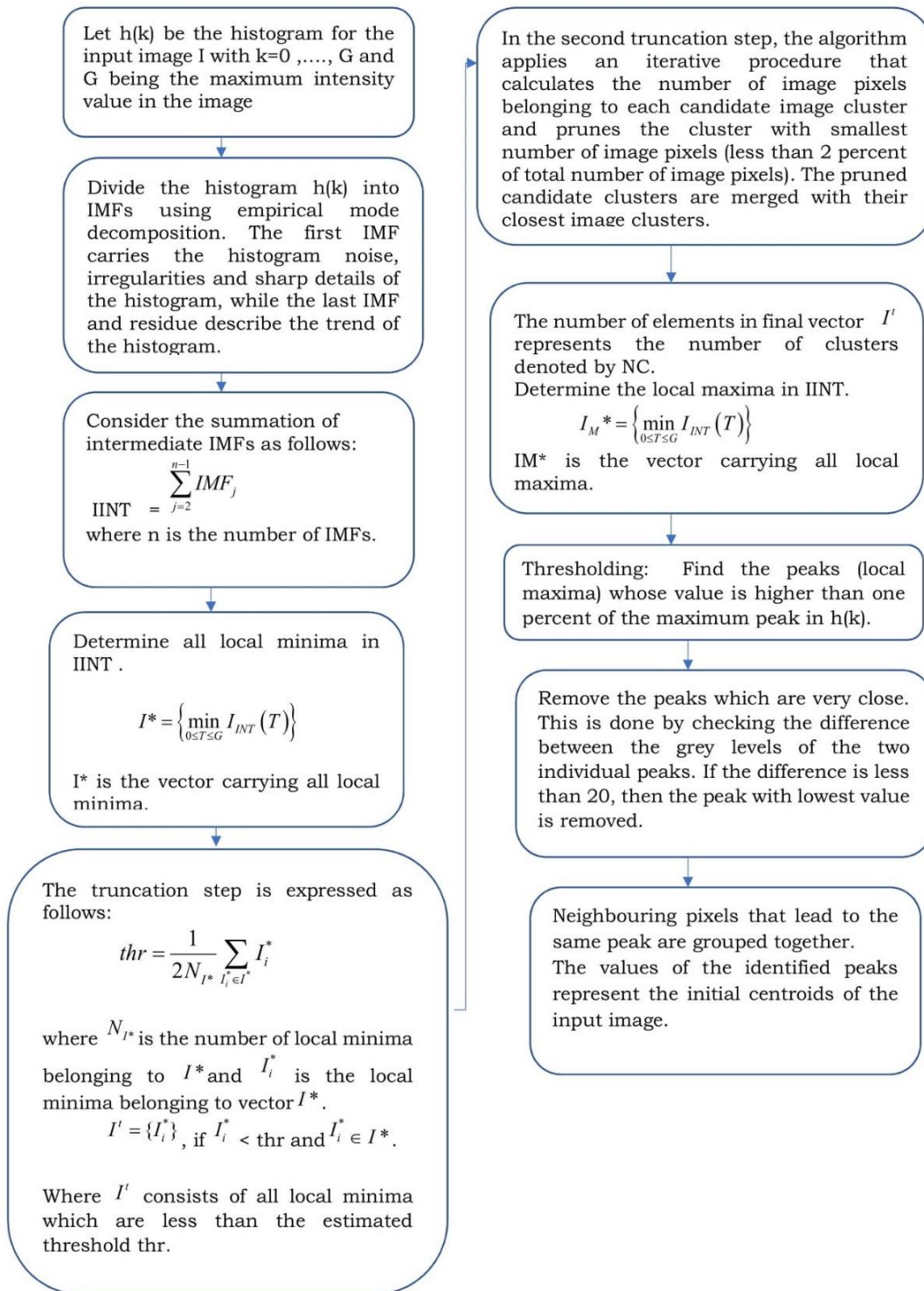


Figure 3.2: Algorithm for estimating clustering parameters

Microarray pictures are segmented using k-mean, moving k-mean, and FCM algorithm. Estimates are made for the two parameters needed by these methods.

4. RESULTS AND DISCUSSIONS

On two microarray pictures of breast cancer as well as a CHG tumour tissue, the proposed sub-gridding and spot-recognition algorithms are evaluated qualitatively and quantitatively. Results are presented in this paper. Figure 6.5 depicts a qualitative study of the suggested sub-algorithm and its components. Figure 6.5 depicts a qualitative examination of the suggested spot identification method based on the data collected. As shown in Table 6.1, a quantitative comparison of the suggested mesh generation algorithms with respect to current approaches has been conducted.

The accuracy of the mesh creation method may be calculated using the formula below.

$$\text{Percentage accuracy} = \frac{\text{Number of spots perfectly gridded}}{\text{Total number of spots}} * 100$$

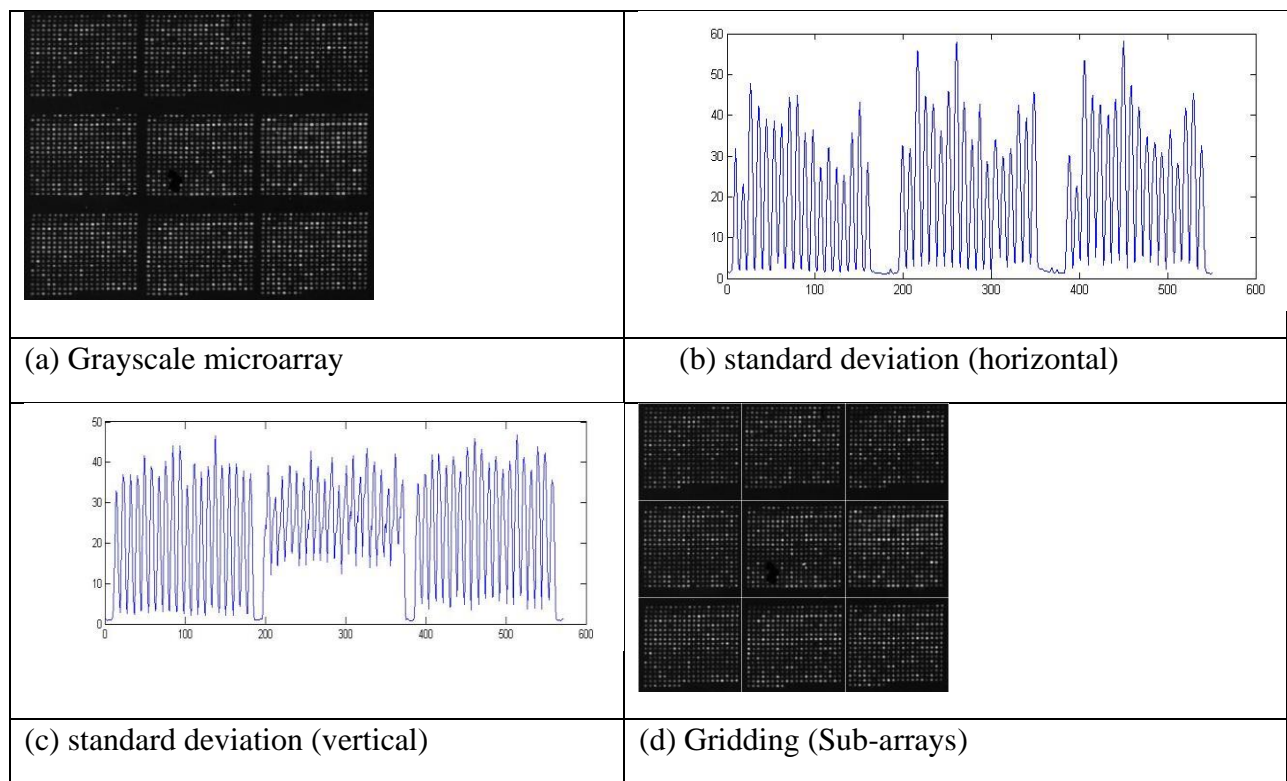
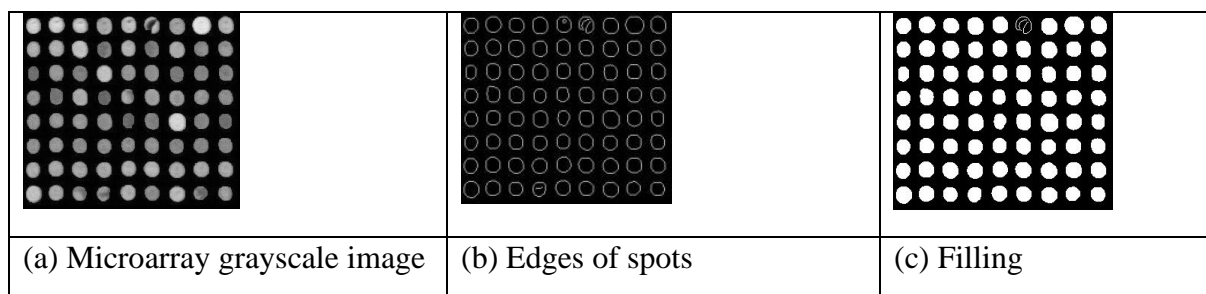


Figure 6.4: Sub-gridding



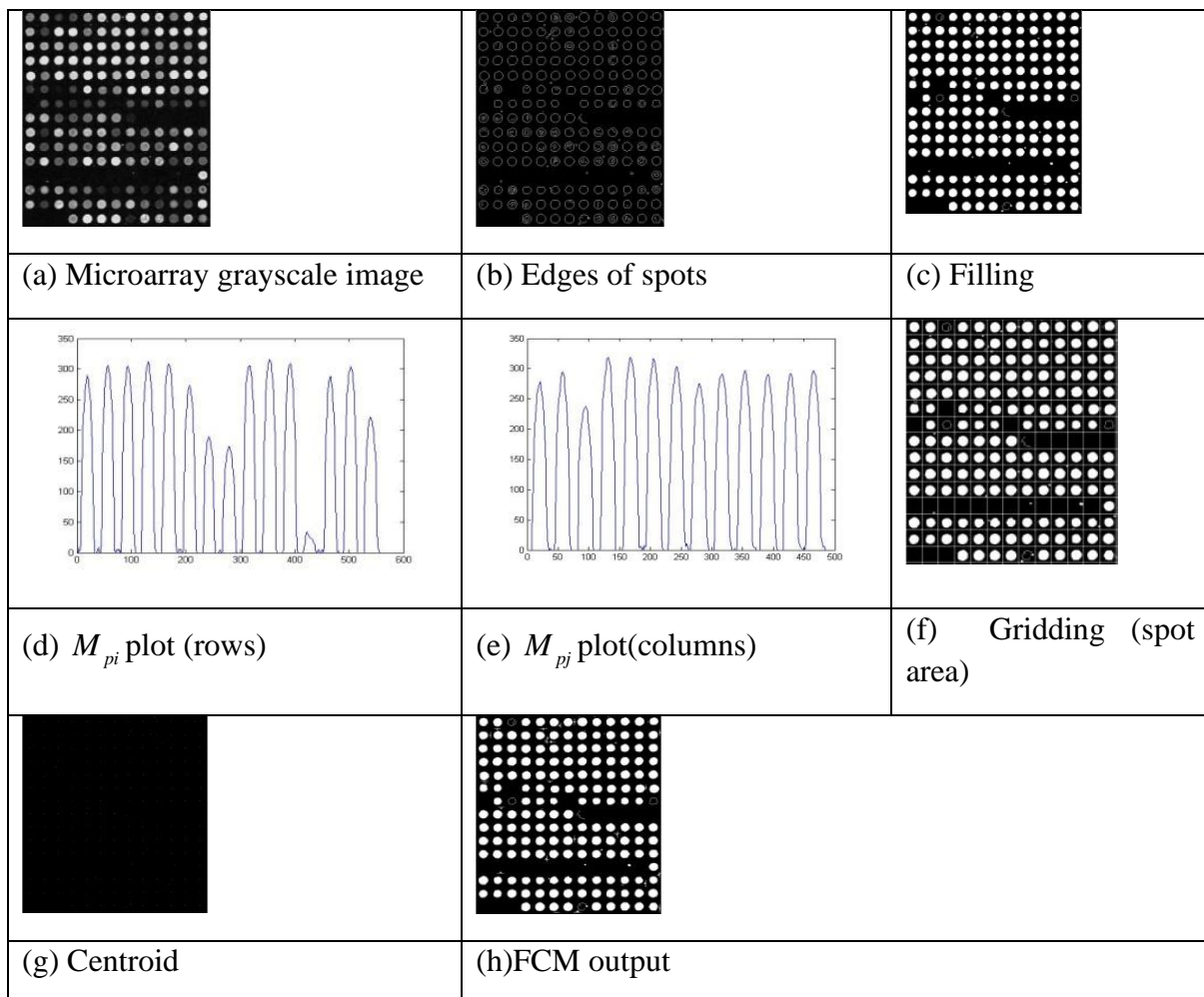
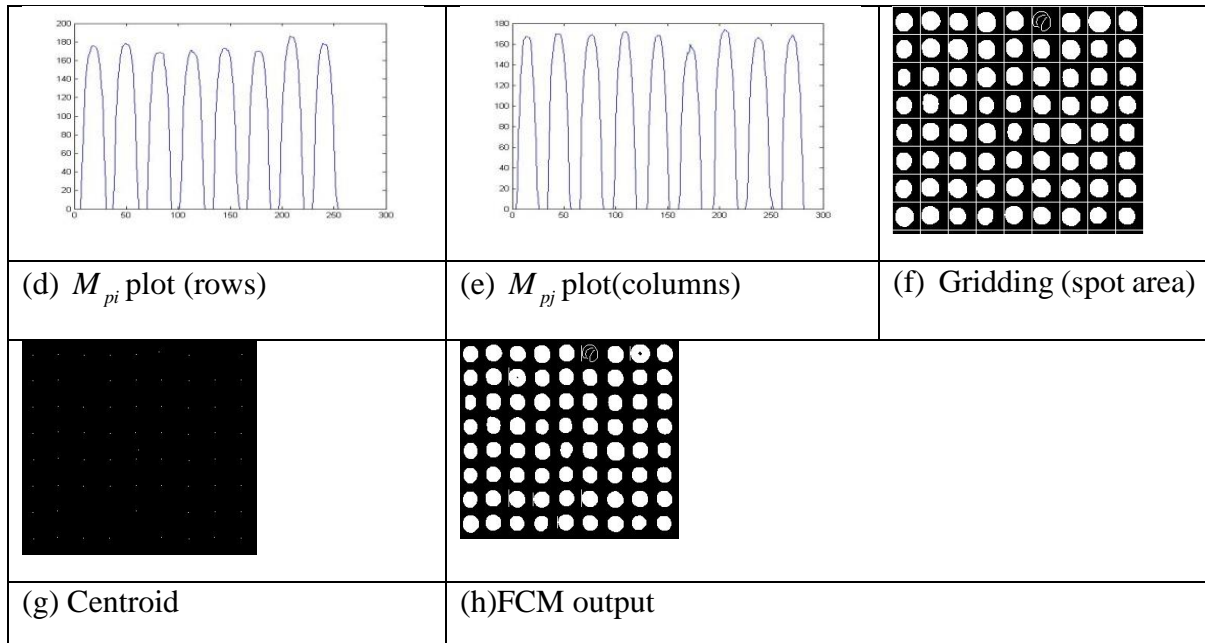


Figure 4.1: Spot-Detection

Table 4.1: Percentage Accuracy of Gridding

Method	Percentage Accuracy Image 1	Percentage Accuracy Image 2
Hirata [Microarray Gridding by mathematical Morphology]	87	79
Jain [Fully Automatic Quantification of microarray image data]	89	81
Wang [Precise Gridding of microarray images by detecting and correcting rotations in subarrays]	91	82
Shuqing Zhao [Microarray Image Processing Based on Mathematical Morphology]	90	84
Deepa .J [A New Gridding Technique for High Density Microarray Images Using Intensity Projection Profile of Best Sub Image]	92	88
Proposed	96	91

Table 4.2: Comparison of iterative steps

	Clustering algorithm	Iterative steps (without ECNC)	Iterative steps (with ECNC)
(Compartment No 1) In image 1	K-means	10	4
	Moving k-mens	14	6
	Fuzzy C-means	17	9

	Clustering algorithm	Iterative steps (without ECNC)	Iterative steps (with ECNC)
(Compartment No 8) In image 2	K-means	11	6
	Moving k-mens	16	12
	Fuzzy C-means	19	11

Table 4.3: MSE values

Method	MSE Values (Compartment No 1) In image 1	MSE Values (Compartment No 8) In image 2
K-means	282.781	346.47
Moving K-means	226.92	238.79
Fuzzy C-means	208.327	196.276

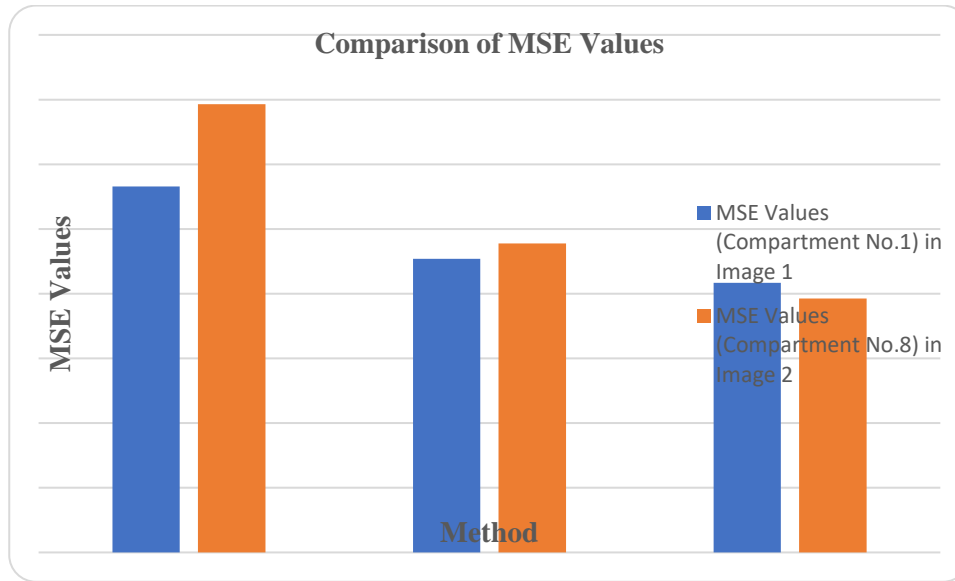


Figure 4.3: Comparatative Results

Individual spots are retrieved after sub-gridding and spot identification methods have been used. They were segmented using a clustering approach that uses a centriods algorithm, as well as an algorithm for estimating how many groups there are in a cluster, to begin with. A clustering method is used to extract a sample location from the image and segment it, which is described in detail below. A total of three distinct clustering methods are employed in the study that is being presented. With and without estimate of starting parameters, the number of iterations utilised by various clustering methods for segmentation of the spot picture is shown in Table 6.2 for each algorithm. Various segmentation strategies are shown in Table 3 by their mean square error (MSE) values [11].

Figure 4.5 depicts a segmented step performed on a single point in time.

Image 1	Gridding	Spot area2
Histogram	IMF1	IMF2

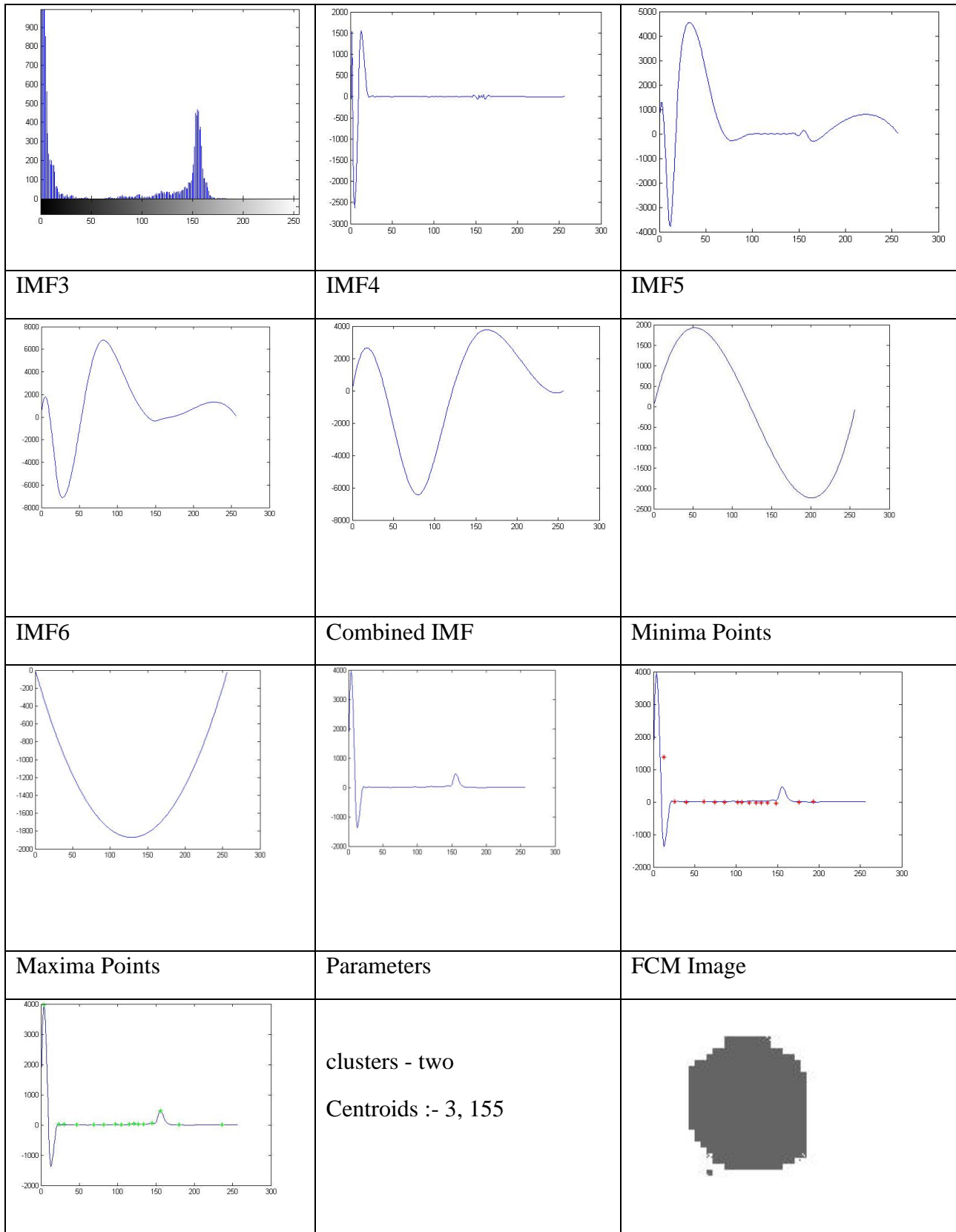
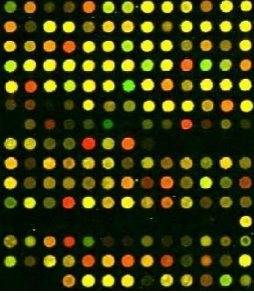
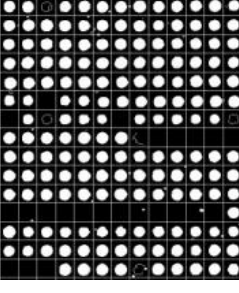
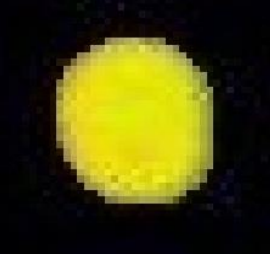
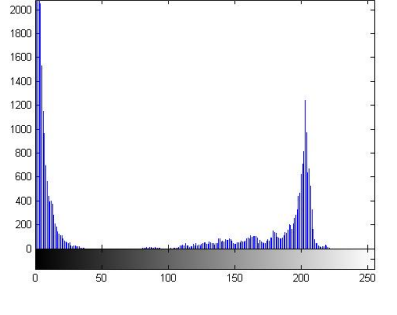
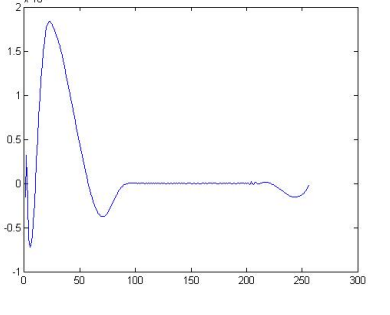
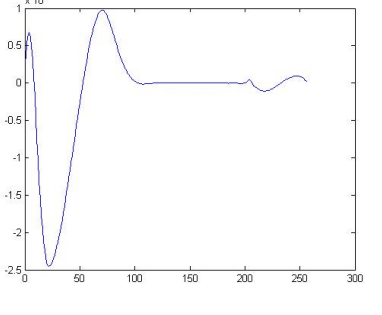
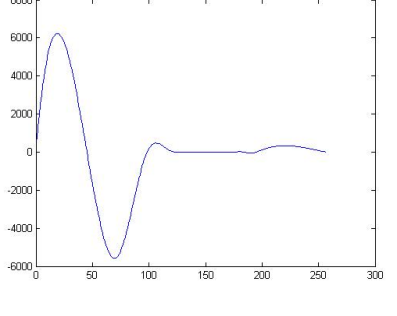
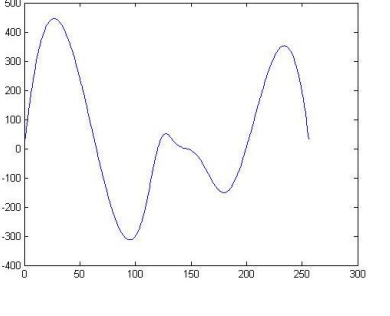
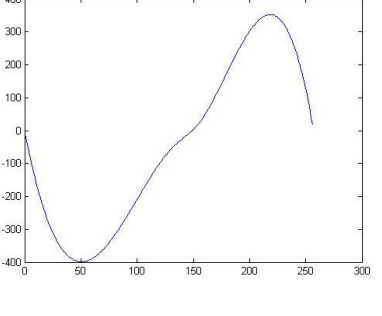
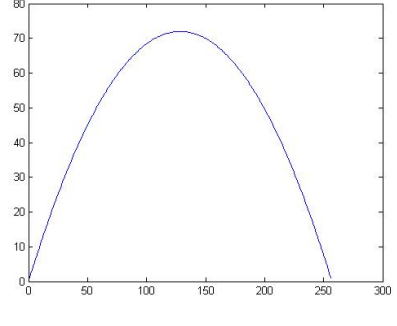
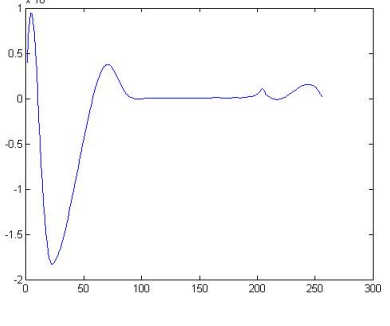
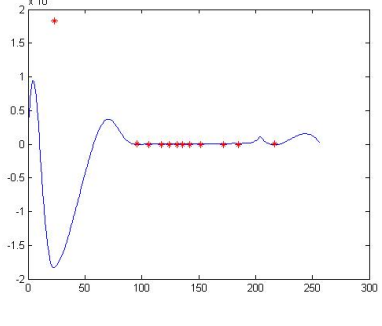


Image 2	Gridding	Spot area7
		
Histogram	IMF1	IMF2
		
IMF3	IMF4	IMF5
		
IMF6	Combined IMF	Minima points
		
Maxima Points	Parameters	FCM Image

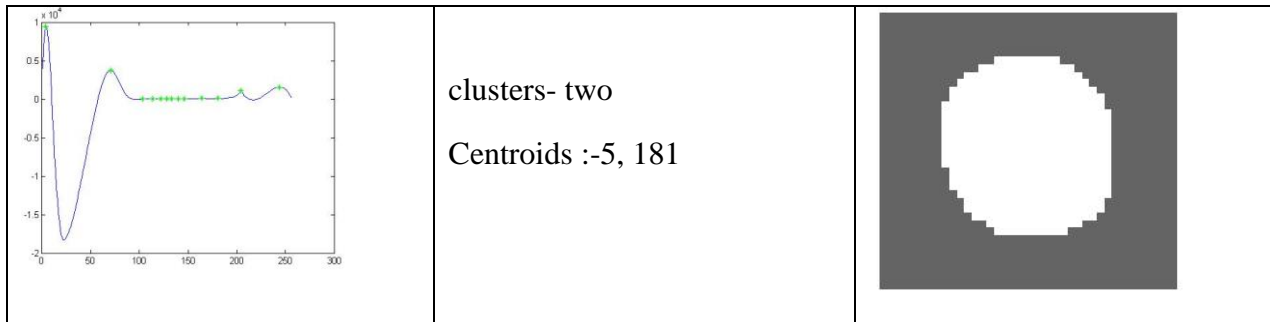


Figure 4.6: Segmented Result

5. CONCLUSION

By employing the IG algorithm, the proposed feature selection generates a ranking of features according to their weight values, which results in a subset of features to be ranked. Later, each unique subset was subjected to BA algorithms and then processed, yielding the best features for further categorization and further processing. Also noted in this categorization subset of factors that have an influence on the decrease of FPR is the fact that In addition, we may do research on multi-classification systems. Any errors made during the gridding and segmentation steps will have an impact on the gene expression value. A gridding algorithm for segmentation is presented in this Paper, as well as an estimation of the parameters required for clustering in the segmentation process. As demonstrated by the experimental results, the proposed algorithm grids an image with 96 percent accuracy while also reduce its overall number of iterations through the estimation of necessary parameters for segmenting the image by a clustering algorithm. A gridding algorithm for segmentation is presented in this Paper, as well as an estimation of the parameters required for clustering. As shown by the experimental findings, the suggested technique grids a picture with 96 percent accuracy while also are reducing its overall number of iterations via the estimation of necessary parameters for segmenting the image by a clustering algorithm.

REFERENCES

1. J.Harikiran et. al., “Fast Clustering Algorithms for Segmentation of Microarray Images”, IJSER, Volume 5, Issue 10, pp 569-574, October 2014.
2. J.Harikiran et. al., “Noise Removal in Microarray Images using Variational Mode Decomposition Technique”, TELKOMNICA Telecommunication Computing Electronics and Control, Vol. 15, No. 4, pp. 1750 – 1757, December 2017.
3. J.Harikiran et. al., “Spot Edge detection in Microarray Images using Bi-dimensional Empirical Mode Decomposition”, Proceedings of C3IT-2012, Vol 4, pp 19-25.
4. M.Katzer at.al.,” Methods for automatic microarray image segmentation” IEEE transactions on NanoBioscience, vol 2, No 4, pp. 202-214, 2003.
5. L.Reuda et.al., “A pattern classification approach to DNA image segmentation”, Pattern Recognition in Bioinformatics, lecture Notes in Computer Science, Volume 5780, pp, 319-330, 2009.
6. J.Rahnenfuhreret.al.,”Hybrid Clustering for microarray image analysis combining intensity and shape features”, BMC bioinformatics, volume 5, No 47, pp.1-11, 2004.

7. N.Karimiet.al.,”Segmentation of DNA microarray images using an adaptive graph based method”, IET Image Processing Vol 4, No1 pp.19-27, 2010.
8. E.Zacharia et.al.,”3D spot modelling for automatic segmentation of cDNA microarray images”, IEEE Transactions on NanoBioscience, Volume 9, pp. 181-192, November 2010.
9. B.Sivalakshmi et.al.,”Microarray Image Analysis using k-means Clustering algorithm”, IJRAT, Vol 6, No 12, December 2018.
10. B.Sivalakshmi et.al, “Microarray Image Analysis using Adaptive Data Clustering Algorithms”, IJAER, Vol 14, No 7, pp.1638-1643, 2019.
11. J. Lyons-Weiler, S. Patel, and S. Bhattacharya. A classification-based machine learning approach for the analysis of genome-wide expression data. *Genome Research*, 13(3):503–512, 2003.
12. H. H. Milioli, R. Vimieiro, C. Riveros, I. Tishchenko, R. Berretta, and P. Moscato. The discovery of novel biomarkers improves breast cancer intrinsic subtype prediction and reconciles the labels in the metabarc data set. *PLOS One*, 10(7):e0129711, 2015.
13. Osmar, R. and Zaiane. “Principles of Knowledge Discovery in Databases - Paper 8 Data Clustering” and Shantanu Godbole data mining Data mining Workshop. 2003.
14. Williams, G., Baxter, R., He, H., Hawkins, S. and Gu, L. “A Comparative Study for RNN for Outlier Detection in Data Mining”, In Proceedings of the 2nd IEEE International Conference on Data Mining, Maebashi City, Japan, pp.709, 2002.
15. Bradley, P. S. and Fayyad, U.M. “Refining Initial Points for K-Means Clustering”, ACM, Proceedings of the 15th International Conference on Machine Learning, pp. 91-99, 1998.
16. Jaing, M., Tseng, S. and Su, C. “Two-phase Clustering Process for Outlier Detection”, *Pattern Recognition Letters*, Volume 22, pp: 691-700, 2001.
17. Kuo, R. J., Liao, J. L. and Tu, C. “Integration of ART2 neural network and genetic Kmeans algorithm for analyzing Web browsing paths in electronic commerce”, *Decision Support Systems*, Vol 40, pp: 355-374, 2005.
18. Yinghua Zhou., Hong Yu. and Xuemei Ca. “A Novel k-Means Algorithm for Clustering and Outlier Detection”, Second International Conference on Future Information Technology and Management Engineering (FITME '09), 476–480, 2009.
19. Xuhui Chen. and Yong Xu. “K-Means Clustering Algorithm with Refined Initial Center”, 2nd International Conference on Biomedical Engineering and Informatics (BMEI '09), pp. 1–4, 2009.
20. Alhadidi, H. N. Fakhouri, and O. S. Al Mousa, “cDNA Microarray Genome Images Processing Using Fixed Spot Position”, *American Journal of Applied Sciences* volume :3, pp 1730-1734, 2006.
21. Bozinov, D., &Rahmenführer, J. Unsupervised Technique for Robust Target Separation and Analysis of DNA Microarray Spots through Adaptive Pixel Clustering. *Bioinformatics*, 18(5), 747-756, 2002.
22. M. T. Miller, A. K. Jerebko, J. D. Malley, and R. M. Summers, “Feature selection for computer-aided polyp detection using genetic algorithms,” in *Proc. of SPIE*, Santa Clara, pp. 102-110, 2003.

23. P. O'Neill and G.D. Magoulas, "Improved processing of microarray data using image reconstruction techniques", IEEE Transactions on Nano bioscience. Volume 2, pp 176–183, 2003.